Tesis Doctoral

MEJORA de la CALIDAD y la PARAMETRIZACIÓN de la VOZ ESOFÁGICA

Presentada por Don Ibon Oleagordia Ruiz

dentro del Programa de Doctorado

CIENCIAS DE LA COMPUTACIÓN

Dirigida por la Dra. Dña. Mª Begoña García Zapirain

Bilbao, 20 de enero de 2015

Aitziberri eta nire familiari

Resumen

Se presenta esta tesis con el propósito de ayudar a las personas que han sufrido una laringectomía, es decir, a las personas laringectomizadas. Con este trabajo de investigación se pretende dar solución a la problemática que, en cuanto a la comunicación, sufren estas personas y, a la monitorización de la voz en el proceso de aprender a hablar con voz esofágica. Con este objetivo y, con el de realizar una aportación a la comunidad científica internacional en el mundo de la investigación, se presentan sendos algoritmos de manera que uno de ellos mejore la calidad de la voz y, que el otro evalúe dicha mejora, midiendo de forma objetiva y automatizada los principales parámetros acústicos de la voz. Es decir, con el primer algoritmo se mejora la inteligibilidad de los laringectomizados en las comunicaciones y, con el segundo se objetiva dicha mejora, además de servir como herramienta para la monitorización de dicho colectivo en el proceso de aprender a hablar con voz esofágica.

Los principales parámetros tratados en el algoritmo de mejora de la voz han sido el Shimmer y Harmonic to Noise Ratio (HNR, Relación armónico ruido). En este procesado se han utilizado técnicas de transformada wavelet, filtrado de Kalman y estabilización de polos. Además de los dos parámetros mencionados, en el algoritmo de parametrización de la voz se han medido los parámetros Pitch y Jitter.

Los resultados demuestran que se ha producido una mejora en el Shimmer de 0,576 dB de media en las 30 voces de la base de datos, quedando este parámetro en los rangos de normalidad de las voces sanas. El parámetro HNR ha experimentado una mejora de 3,459 dB de media. Subjetivamente, el algoritmo propuesto reduce sustancialmente el ruido de aspiración del esófago. En cuanto a la parametrización de la voz, se observa que el algoritmo propuesto es claramente mejor medidor que el Gold Standard para las voces esofágicas y que es tan buen medidor para las voces sanas e incluso mejor en ciertos parámetros.

Abstract

This thesis is presented with the purpose of helping people who have experienced a laryngectomy, i.e. laryngectomees. The aim of this research work is to provide a solution to the problem suffered by these people regarding communication. A solution is also provided to monitoring of the voice in the process of learning how to speak with an esophageal voice. In fulfilling this objective, together with the aim of making a contribution to the international scientific community in the world of research, two algorithms are presented, insofar as one of them improves the quality of voice, and the other one evaluates such improvement, thus measuring the main acoustic parameters of the voice in an automated and objective manner. In other words, with the first algorithm, the intelligibility of laryngectomees improves in communication, and, with the second one, such improvement is objectified, in addition to serving as a tool for monitoring this group in the process of learning how to speak with an esophageal voice.

The main parameters treated in the voice improvement algorithm were shimmer, and the Harmonic to Noise Ratio (HNR). Wavelet transform, Kalman filter and pole-stabilization techniques were used in this process. Aside from the aforementioned parameters, in the voice parameterization algorithm, Pitch and Jitter parameters were also measured.

The results show an improvement of an average 0.576 dB in the case of the Shimmer, for the 30 voices in the database, this being parameter within the ranges of normality for healthy voices. The HNR parameter experienced an improvement of 3.459 dB. Subjectively, the esophageal breathing noise is reduced substantially. As far as voice parameterization is concerned, the proposed algorithm can be clearly seen to be a better gauge than the Gold Standard for esophageal voices, and is a good gauge for healthy voices, or even better in certain parameters.

Laburpena

Tesi hau laringektomia kirurgia izan duten pertsonei laguntzeko aurkezten da, hau da, pertsona larigektomizatuak. Ikerketa-lan honek, alde batetik, pertsona hauek komunikatzeko daukaten arazoari soluzioa eman nahi dio, eta bestetik, esofagoko ahotsarekin hitz egiten ikasteko prozesuan ahots hori monitorizatzeko balioko du. Helburu honekin eta komunitate zientifikoan ekarpen berria egiteko asmoz, algoritmo bi aurkezten dira. Batak, esofagoko ahotsaren kalitatea hobetzen du, eta besteak, hobekuntza hori ebaluatzen du ahotsaren parametro esanguratsuenak era objektibo eta automatikoan neurtuz. Hau da, lehenengo algoritmoak komunikatzerakoan laringektomizatuen ulergarritasuna hobetzen du, eta bigarrenak, hobekuntza hori neurtu egiten du. Era berean, kolektibo honek esofagoko ahotsarekin hitz egiten ikasteko prozesuan ahots horren kalitatea neurtzeko eta monitorizatzeko tresna bat da.

Ahotsaren kalitatea hobetzeko Shimmer eta Harmonic to Noise Ratio (HNR, armoniko-zarata ratioa) parametroak jorratu dira. Prozedura honetan wavelet transformatua, Kalman iragazkia eta poloen egonkortze teknikak erabili dira. Aipatutako bi parametroez aparte, parametrizaziozko algoritmoak beste bi neurtzen ditu, Pitch eta Jitter, alegia.

Lortutako emaitzek frogatzen dute Shimmerak 0,576 dB-ko hobetu duela batez bestez datu-basearen 30 ahotsetan. Honela, parametro horrek ahots osasungarrien mailan lortu du. HNR parametroak 3,459 dB-ko hobekuntza lortu du eta batez beste. Subjektiboki, arnastearen zarata ezabatu egin dela esan daiteke.

Parametrizazioari dagokionez, argi ikus daiteke proposatutako algoritmoa Gold standard-a baino neurtzaile hobeagoa dela esofagoko ahotsetan. Eta Gold standard-a bezain ona edo hobeago hainbat kasutan ahots osasungarrietan.

Agradecimientos

Quisiera dar mi más sincero agradecimiento a aquellas personas que han colaborado o contribuido en mayor o menor medida en este trabajo de investigación.

En primer lugar me gustaría dar las gracias a mi directora de tesis, Begoña García Zapirain, por estar ahí cuando era necesario, por sus consejos, por su paciencia, por su energía… por todo.

Me gustaría dar las gracias también a todos los compañeros de Deustotech-Life y a todos aquellos que han estado colaborando con esta investigación. Merecen especial mención Mikel M., Andrés G. y Alejandro B.

Agradecer también la colaboración de la Asociación Vizcaína de Laringectomizados, por aportar las voces que han hecho posible completar la base de datos.

Por último, me gustaría dar las gracias de manera muy especial a toda mi familia y sobre todo a Aitziber, que siempre ha estado a mi lado apoyándome, dándome ánimos y buenos consejos.

Muchas gracias a todos por estar ahí, por colaborar, por apoyar, por alentar, por estar a mi lado cuando hacía falta. Gracias a todos de corazón.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN

1. INTRODUCCIÓN

"Tanta prisa tenemos por hacer, escribir y dejar oír nuestra voz en el silencio de la eternidad, que olvidamos lo único realmente importante: vivir"

(Robert Louis Stevenson)

Las personas que se les ha extirpado totalmente la laringe debido a un cáncer se les llama laringectomizados. La extracción de la laringe, o laringectomía, ha demostrado a lo largo de los años ser un método muy eficaz para el tratamiento de los tumores de la laringe. Esta intervención se propone ante una enfermedad grave como puede ser el cáncer de laringe.

El cáncer de laringe es el segundo cáncer en incidencia del tracto aerodigestivo superior (95%), siendo el carcinoma escamoso el tipo histopatológico el predominante. En un año se diagnostican aproximadamente 136.000 nuevos casos de cáncer de laringe en el mundo, con una supervivencia global de 5 años, un 68% [Tirado+07]. Desde que Theodore Billroth realizó en 1873 la primera laringectomía total con éxito, uno de los objetivos de los otorrinolaringólogos ha sido no sólo velar por la supervivencia de los pacientes, sino procurarles una calidad de vida aceptable. La laringectomía total es una cirugía mutilante y aunque la calidad de vida es razonablemente buena, una de las mayores discapacidades a las que debe enfrentarse el sujeto es la pérdida de la voz [Puo04] [Vazquez+05] [Weymuller+00].

Una de las consecuencias de la laringectomía es la pérdida del habla del paciente (o voz laringada). Una vez recuperado puede volver a hablar mediante una prótesis fonatoria o bien mediante el aprendizaje del habla erigmofónica o habla esofágica, aunque no siempre es posible su rehabilitación. Cuando la recuperación de la voz es posible, a la voz emitida por los larigectomizados se le llama **voz esofágica**. Tres son básicamente las opciones con las que cuenta el laringectomizado para restituir su capacidad de comunicación oral: voz esofágica o erigmofónica, voz esofágica con prótesis fonatoria y la electrolaringe.

Otras consecuencias de la operación son: la necesidad de respirar mediante un traqueostoma, abierto en el cuello para tal fin, la disminución en el sentido del olfato, la disminución de la fuerza para levantar pesos y las dificultades en la contracción del abdomen para hacer de vientre u orinar debido a que no es posible contener la respiración.

La operación consiste en extirpar toda la laringe dañada por un tumor. Si se realiza una laringectomía total, la laringe se extirpa completamente lo que implica la pérdida de la voz y la necesidad de llevar permanentemente un traqueostoma (orificio en el cuello) para poder respirar. En esta operación se crea una nueva abertura, llamada estoma, mediante la cual se efectúa la respiración. A veces, más adelante, y tras un aprendizaje, o mediante una prótesis fonatoria externa o mediante una técnica operatoria especial llamada fístula fonatoria, puede llegar a hablar de nuevo.

Después de la operación, durante la rehabilitación, el paciente comenzará el proceso de aprendizaje para emitir voz esofágica, es decir, la voz producida por la modulación del aire que proviene del esófago. La voz esofágica o erigmofónica exige un aprendizaje más o menos arduo que requiere frecuentes explicaciones para que el sujeto comprenda el principio fundamental en el que se basa su producción: inyección, succión o deglución de aire (elemento efector) desde la cavidad oral hacia el segmento faringoesofágico, para provocar una erupción automática fluida o en su defecto una eructación voluntaria del flujo ascendente

de aire hacia la cavidad oral. Es este flujo de aire la fuente de energía necesaria para producir la vibración de la mucosa redundante de la neoglotis (elemento vibrador), originando la frecuencia fundamental que posteriormente se enriquece en segmentos más superiores del tracto aerodigestivo (elemento resonador) para proyectarse en forma de voz inteligible mediante la articulación de la palabra (elemento articulador) [Kearney04]. Este hecho permite comunicarse al paciente, eso sí, con gran dificultad para mantener conversaciones fluidas debido a la baja calidad de la voz esofágica.

Con el propósito de ayudar a las personas que han sufrido una laringectomía se presenta esta investigación. Para ello, por un lado, diseñaremos algoritmos de procesado digital de señal de cara a que la voz esofágica se parezca más a la voz laringada, y por el otro lado, realizaremos algoritmos para la parametrización de la voz. Con esta tesis nos proponemos dar solución a la problemática que sufren las personas laringectomizadas en cuanto a la comunicación y a la monitorización de la voz en el proceso de rehabilitación. Se intentará mejorar la calidad de la voz, mejorando así la inteligibilidad en la comunicación. De cara a evaluar dicha mejora, será necesario medir de forma objetiva los principales parámetros acústicos de la voz.

Por lo tanto, para que la voz esofágica se parezca más a la voz laringada, en esta tesis hemos diseñado algoritmos de cara a mejorar los parámetros acústicos HNR y Shimmer de la voz esofágica. Nuestros algoritmos están centrados en los citados parámetros ya que son unos de los más relevantes en la inteligibilidad de la voz y están validados por la comunidad científica internacional.

Por otro lado, caracterizaremos uno de los principales parámetros acústicos de la voz (tanto para la voz laringada como para la esofágica), como es el "Pitch" o frecuencia fundamental, del cual se pueden obtener otros de especial relevancia en la inteligibilidad de la voz y en la monitorización en el proceso de rehabilitación como pueden ser: el "Jitter" o variación de la frecuencia fundamental, "Shimmer" o variación de la amplitud de la frecuencia

fundamental y la relación armónico-ruido (Harmonic to Noise Ratio, "HNR" en adelante). Dicha caracterización se realizará de forma automática para la voz esofágica.

Para evaluar la mejora de la calidad de la voz esofágica y de cara a comparar las mediciones de los parámetros acústicos con un paquete de software comercial, se utilizará "Multi Dimensional Voice Program", en adelante MDVP [Deliyeski93] [Nicastri+04]. Ha sido necesario realizar estas mediciones a mano, indicando en la aplicación dónde se encuentran los instantes de Pitch o los ciclos o épocas de la voz ya que, como se ha mencionado anteriormente, el paquete de software comercial no mide de forma correcta y automática la voz esofágica.

Sin embargo, unos de los mayores problemas es que no se puede evaluar de forma objetiva este tipo de voz esofágica durante el proceso de rehabilitación, ya que no existe en el mercado ninguna aplicación que pueda obtener los parámetros acústicos automáticamente para este tipo de voz. La calidad de la voz esofágica es tan baja que los paquetes de software comerciales y los algoritmos que obtienen la periodicidad de la voz o Pitch no funcionan adecuadamente y, por lo tanto, las medidas que obtienen dichos paquetes de software no son fidedignas.

Para solucionar la problemática de la caracterización (o valoración objetiva) de la voz esofágica y de su baja inteligibilidad, esta tesis propone técnicas de procesado digital de señal medir los principales parámetros de la voz.

Hasta hace relativamente poco tiempo la valoración objetiva y exacta de la voz era inexistente. Muchos otorrinolaringólogos opinaban que los únicos instrumentos que se requerían para el estudio de la voz son el propio oído y los espejos laríngeos y que la llamada valoración objetiva y los novedosos "juguetes informáticos" resultaban innecesarios. Sin embargo, los tiempos cambian y como cualquier función humana, la voz puede padecer alteraciones y su patología debe ser investigada. Al mismo tiempo esta investigación tiende a la objetivación de la alteración fonatoria con tres fines principales:

a) para comparar resultados post-tratamiento (médico, de logopedía o quirúrgico) y evolución de la rehabilitación.

b) y cada vez más frecuente, para solucionar conflictos médico-legales.

La valoración objetiva de la voz se puede realizar mediante el estudio morfo-funcional (exploración física del órgano fonatorio con la laringoestroboscopia) y el análisis acústico vocal (estudio de los principales parámetros acústicos que componen la voz humana).

1.1 HIPÓTESIS

Con el fin de delimitar con mayor precisión el ámbito del trabajo a realizar, presentamos en este apartado la hipótesis cuya verificación promueve la presente investigación y los objetivos fijados para la tesis, dividiendo éstos en generales y específicos.

La hipótesis cuya verificación promueve la presente investigación puede resumirse de forma siguiente:

Es posible mejorar la calidad de la voz esofágica y caracterizar dicha voz de forma automática utilizando algoritmos de procesado de señal.

En cuanto a la mejora de la calidad de la voz esofágica, se realizarán algoritmos que mejoren los parámetros acústicos de la voz esofágica de tal manera que se asemejen a la voz laringada. En estos algoritmos de mejora nos centraremos especialmente en dos parámetros acústicos, como ya se ha mencionado anteriormente: la relación armónico-ruido, HNR, y el Shimmer.

Por otro lado, la caracterización de la voz se realizará un algoritmo que detecte automáticamente los instantes de Pitch para todo tipo de voces. Hemos realizado una clasificación de voces en dos grupos: la voz laringada o sana y la voz esofágica. Una vez detectados los instantes de Pitch para los distintos grupos de voces, se calcularán los parámetros acústicos citados validados por la comunidad científica internacional: Pitch o frecuencia fundamental, Jitter, Shimmer y HNR.

1.2 OBJETIVOS

De cara a la demostración de la validez de dicha hipótesis, fijamos a continuación el objetivo general del presente trabajo de investigación así como los objetivos específicos que deberán llevar a la consecución del primero.

1.2.1 Objetivo General

El principal objetivo de esta tesis se podría enunciar de la siguiente manera:

El objetivo principal es mejorar la calidad de la voz de los laringectomizados y realizar la caracterización de dicha voz para que puedan monitorizar su evolución en el proceso de aprendizaje.

Durante la rehabilitación, los laringectomizados pueden conseguir el habla con bastante facilidad aunque no sin trabajo y constancia. Los altibajos son constantes entre el colectivo. Es decir, hay días que todo se ve de color oscuro y otros, por el contrario, todo se ve más claro, cuando las cosas salen mejor. El progreso del habla durante la rehabilitación es muy importante para su ánimo y autoestima. El hecho de que haya una herramienta, dentro del campo de las Tecnologías de la Información y Comunicación (TIC), que les ayude a mejorar la calidad de la voz esofágica hace que su ánimo aumente y mejoren en el proceso de aprendizaje. De esta manera, las personas que ya dominan el habla esofágica les sirve de estímulo a las personas que están aprendiendo. Todo ello revierte en la calidad de vida de las personas laringectomizadas.

1.2.2 Objetivos Técnicos de Mejora de la Voz Esofágica

Hay un grupo de objetivos de carácter técnico relacionados con la mejora de la voz esofágica y son las siguientes:

- ***Diseñar un algoritmo que modifique el espectro-temporal de la señal de voz mediante técnicas de la Transformada Wavelet.***
- ***Diseñar un nuevo algoritmo que disminuya el ruido de la señal de voz mediante técnicas de Filtrado de Kalman.***
- ***Concatenar estos dos algoritmos con el ya existente de estabilización de polos*** **[García03].**

En lo que respeta a los objetivos técnicos, a continuación se detalla qué se persigue con cada uno de ellos.

El primero de los objetivos de carácter más científico es el de "***Diseñar un algoritmo que modifique el espectro-temporal de la señal de voz mediante técnicas de la Transformada Wavelet***". Dentro de la transformada wavelet, hemos utilizado para el diseño de este algoritmo la transformada wavelet discreta (Discrete Wavelet Transform, DWT). Para ello, se planteará la mejora de algunas características de la voz, especialmente el *Shimmer* o variación de la amplitud en los instantes de Pitch y la relación armónico-ruido (*HNR*).

Al mismo tiempo y, durante el transcurso de esta investigación, se ve absolutamente necesario reforzar la mejora del ruido existente en la señal de voz esofágica. Con este objetivo se plantea el "***Diseñar un nuevo algoritmo que disminuya el ruido de la señal de voz mediante técnicas de Filtrado de Kalman***". Se ha adaptado este algoritmo aprovechando las características propias de la señal de voz esofágica, concretamente, se ha aprovechado el ruido de los momentos de silencio.

Una de las técnicas que se han empleado en anteriores investigaciones, como se ve en la tesis doctoral de Begoña García Zapirain [García03], son la estabilización

o modificación de los polos y los ceros del sistema que modeliza la voz mediante técnicas de procesado de los **Linear Prediction Coeficient (LPC)**. En esta tesis hemos concatenado los tres algoritmos para una mayor optimización de la señal de voz esofágica.

1.2.3 Objetivos Técnicos de Caracterización de Señal de Voz

El objetivo técnico relacionado con la caracterización de la señal de voz esofágica es:

- ***Desarrollar un algoritmo que permita cuantificar objetivamente los parámetros de la voz esofágica, diferenciando distintas calidades de voz.***

En relación al objetivo de "***Desarrollar un algoritmo que permita cuantificar objetivamente los parámetros de la voz esofágica, diferenciando distintas calidades de voz***", esta investigación se ha centrado en medir el Pitch con exactitud y para distintas calidades de voz. Las distintas voces utilizadas son: la voz laringada o voz sana y la voz esofágica. Para la medición del Pitch de una forma correcta y automática, en el algoritmo se ha tenido que distinguir mediante el Pitch las distintas calidades de la voz. Una vez obtenido el Pitch con exactitud, se han podido medir otros parámetros acústicos como son el Jitter, Shimmer y la relación armónico-ruido (HNR).

Estos algoritmos de caracterización de la voz esofágica se han implementado en un paquete software que mide distintos parámetros de señal de voz, cualquiera que sea su calidad o condición, de forma automática.

1.2.4 Objetivos Sociales de la Investigación

El objeto del proyecto paralelo a la realización de esta tesis es el de desarrollar e implantar **paquetes de software de uso**, basadas en **plataformas tecnológicas** para colectivos de discapacitados y personas con enfermedades que les impiden poder tener una autonomía del 100%, promoviendo la **e-inclusión y la e-asistencia de las mismas**.

Por tanto, los objetivos de carácter social de este paquete de software son:

- ***Promover la e-inclusión de las personas laringectomizadas***
- ***Promover la e-asistencia de dicho colectivo***
- ***Potenciar la autonomía global en las comunicaciones telefónicas***

Se pretende poner a disposición del colectivo de laringectomizados, ORL, foniatras y logopedas una herramienta software amigable para la medida de los parámetros acústicos de la voz esofágica en procesos de reeducación del habla que mejoren la efectividad de los métodos existentes.

Este proyecto abarca el siguiente campo de actuación que consisten en ofrecer a las personas discapacitadas la posibilidad de potenciar una **autonomía global** mediante la investigación y la implantación de tecnologías adecuadas. Consiste en investigar y desarrollar un software que asegure que el 100% de los miembros de asociaciones de laringectomizados que estén en procesos de aprendizaje de voz esofágica usen la herramienta al menos una vez durante el mismo.

1.3 METODOLOGÍA DE LA INVESTIGACIÓN

En este apartado se expondrá la metodología llevada a cabo en el proceso de investigación de esta tesis doctoral. Esta metodología está dividida en varias etapas:

- ***Identificación del problema científico***
- ***Hipótesis***
- ***Objetivos para la solución del problema***
- ***Investigación y experimentación***
- ***Validación y pruebas***

- ***Identificación del problema científico***

Por *problema científico* se entiende toda dificultad teórica o práctica que le compete a la ciencia resolver; toda cuestión que se trata de aclarar, las situaciones que no tienen solución conocida; las preguntas que derivan de la observación científica. Es la interrogante que se formula el investigador ante una realidad desconocida o ante la falta de información o información incompleta para explicarse un hecho (laguna o defecto). También surge cuando existen contradicciones o incoherencias en la información científica.

En la introducción de esta tesis se ha abordado con detalle el *problema científico* de esta investigación que no es otro que el de la baja calidad de la voz esofágica y la falta de herramientas o técnicas de procesado señal que caractericen correctamente la misma.

- ***Hipótesis***

En el ámbito científico, la *hipótesis* es la explicación plausible, de tipo racional, de los hechos y fenómenos, y que se acepta provisionalmente con el objeto de someterla a comprobación posterior; admitida como principio que da lugar a un sistema de proposiciones o teoremas, es decir, proposiciones demostrables, las que junto con definiciones constituyen el sistema hipotético−deductivo.

En el caso de este trabajo de investigación el enunciado de la hipótesis es el nexo de unión entre el problema y la solución de la misma. Por lo tanto, se plantea como hipótesis que es posible caracterizar la voz esofágica de forma automática y mejorar dicha voz utilizando algoritmos de procesado de señal.

- ***Objetivos para la solución del problema***

La formulación del o los objetivos de la investigación tiene como finalidad la de expresar y justificar por qué llevamos adelante una investigación. La mayoría de los proyectos de investigación formulan dentro de sus objetivos finales las implicaciones o contribuciones que esperan lograr. Su formulación sirve, por un

lado, para situar al estudio dentro del contexto social general y, por otro, para precisar los resultados que se esperan; lo que viene a reafirmar, al mismo tiempo, su necesidad [Giddens97].

En nuestro caso concreto se han establecido dos tipos de objetivos: el objetivo de carácter más general y los objetivos técnicos.

- ***Investigación y experimentación***

La experimentación es un método común de las ciencias y las tecnologías, consiste en el estudio de un fenómeno, reproducido generalmente en un laboratorio repetidas veces en las condiciones particulares de estudio que interesan, eliminando o introduciendo aquellas variables que puedan influir en él. Se entiende por variable todo aquello que pueda causar cambios en los productos de un experimento y se distingue entre variable único, conjunto o microscópico [Rojas99].

Para llegar a la validación de la hipótesis planteada hemos utilizado el método de *experimentación en el laboratorio* [Straub+04], es decir, una técnica cuantitativa que ha sido aplicada para la validación formal de los resultados. En este caso concreto, se han medido toda la base de datos de voz anteriormente a la aplicación del algoritmo y posteriormente de cara a la validación de la hipótesis.

La *investigación científica* es una actividad orientada a la obtención de nuevos conocimientos y, por esa vía, ocasionalmente dar solución a problemas o interrogantes de carácter científico [Zorrilla07].

Una *investigación* se caracteriza por ser un proceso único:

- **Sistemático**: A partir de la formulación de una hipótesis u objetivo de trabajo, se recogen datos según un plan preestablecido que, una vez analizados e interpretados, modificarán o añadirán nuevos conocimientos a los ya existentes, iniciándose entonces un nuevo ciclo de investigación. La sistemática empleada en una investigación es la del método científico.

- **Organizado**: Todos los miembros de un equipo de investigación deben conocer lo que deben hacer durante todo el estudio, aplicando las mismas definiciones y criterios a todos los participantes y actuando de forma idéntica ante cualquier duda. Para conseguirlo, es imprescindible escribir un protocolo de investigación donde se especifiquen todos los detalles relacionados con el estudio.

- **Objetivo**: Las conclusiones obtenidas del estudio no se basan en impresiones subjetivas, sino en hechos que se han observado y medido, y que en su interpretación se evita cualquier prejuicio que los responsables del estudio pudieran hacer.

Para la elaboración de esta tesis también se utilizado el método de *investigación-acción*, cuya finalidad es tener un enfoque cualitativo que se aplica para la definición, refinamiento continuo y validación de la solución planteada. Concretamente, la investigación-acción es una forma de indagación introspectiva colectiva emprendida por participantes en situaciones sociales con objeto de mejorar la racionalidad y la justicia de sus prácticas sociales o educativas, así como su comprensión de esas prácticas y de las situaciones en que éstas tienen lugar. Se trata de una forma de investigación para enlazar el enfoque experimental de la ciencia social con programas de acción social que respondan a los problemas sociales principales. Dado que los problemas sociales emergen de lo habitual, la *investigación-acción* inicia el cuestionamiento del fenómeno desde lo habitual, transitando sistemáticamente, hasta lo filosófico. Mediante la investigación–acción se pretende tratar de forma simultánea conocimientos y cambios sociales, de manera que se unan la teoría y la práctica [Burns07] [Lewin46].

- **Investigación-acción**

La investigación-acción significa planificar, actuar, observar y reflexionar más cuidadosamente, más sistemáticamente y más rigurosamente de lo que suele hacerse en la vida cotidiana; y significa utilizar las relaciones entre esos

momentos distintos del proceso como fuente tanto de mejora como de conocimiento.

1. **El plan** es de acción organizada y, por definición, debe anticipar la acción: debe mirar hacia delante. Debe reconocer que toda acción social es, hasta cierto punto, impredecible y, en consecuencia, un tanto arriesgada. El plan general debe ser lo bastante flexible para adaptarse a efectos imprevistos y a limitaciones anteriormente indiscernibles. La acción prescrita por el plan debe estar informada críticamente en dos sentidos. En primer lugar, debe tomar en consideración los riesgos que implica un cambio social y reconocer las limitaciones reales, materiales y políticas, de la situación. En segundo lugar, la acción críticamente informada debe ser elegida de tal modo que permita a los profesionales actuar más eficazmente sobre un abanico más amplio de circunstancias, y hacerlo más sabia y prudentemente. Debe ayudar a los profesionales a llegar más allá de las limitaciones actuales (al menos en alguna medida) y capacitarlos para actuar más adecuadamente en la situación dada y resultar más efectivos como educadores. Debe ayudarles a comprender un nuevo potencial para la acción educativa. Deben colaborar, como parte integrante del proceso de planificación, en una discusión (que será un discurso al mismo tiempo teórico y práctico) orientada a formar un lenguaje mediante el cual podrán analizar y mejorar su comprensión y su acción.

   En el caso concreto de esta tesis la planificación se ha llevado a cabo en diferentes ciclos de iteración de la investigación-acción como pueden ser: planificación inicial, búsqueda y análisis de bibliografía, definición de acciones a realizar e hitos, proyectos en los que se va a participar etc.

2. **La acción**, en el sentido en que aquí se entiende, es deliberada y está controlada: es una variación cuidadosa y reflexiva de la práctica, y está informada críticamente. Reconoce en la práctica ideas en acción y utiliza la acción como plataforma para un nuevo desarrollo en la acción posterior,

una acción con un propósito investigador. La acción está guiada por la planificación en el sentido de que mira hacia atrás para planificar su racionalidad. Pero la acción críticamente informada no está completamente controlada por planes. En lo esencial, es arriesgada. Tiene lugar en el tiempo real y se enfrenta a limitaciones materiales reales (algunas de las cuales surgen repentina e impredeciblemente a consecuencia de cambios dentro del marco de acción). En consecuencia, los planes de acción deben poseer siempre una cualidad tentativa y provisional; deben ser flexibles y estar abiertos al cambio, respondiendo a las circunstancias. La acción críticamente informada reconoce también que, en cierta medida, está vinculada a una práctica anterior (aquello que antes se ha hecho, modos previos de actuar); pero la práctica anterior, por su parte, sólo abarca tentativamente las realidades del presente. la acción es, pues, fluida y dinámica y exige decisiones instantáneas, acerca de qué debe hacerse, así como el ejercicio de un raciocinio práctico. La puesta en obra de los planes de acción adquirirá el carácter de una lucha material y social por el logro de la mejora. Quizá sean necesarias la negociación y el compromiso, pero los compromisos deben contemplarse también en su contexto estratégico. Uno de los modos en que la investigación-acción difiere de la acción en las situaciones usuales es que se trata de una acción observada. Los actores intentan recoger datos acerca de su acción con objeto de poder valorarla a fondo.

3. **La observación** es necesaria porque la acción se verá siempre recortada por limitaciones de la realidad, y no siempre se conocerá anticipadamente la existencia de todas esas limitaciones. La observación debe planificarse de tal modo que se constituya una base documental para la reflexión posterior, pero no debe ser demasiado estrecha de miras. La observación, igual que la acción misma, debe ser suficientemente flexible y abierta para registrar lo inesperado. Las personas dedicadas a la investigación-acción deberían registrar siempre en un diario observaciones adicionales a

aquellas que encajan en las categorías planificadas de la observación. La observación anticipa los logros de la reflexión. De ese modo, puede contribuir a mejorar la práctica a través de un grado más alto de comprensión y de una acción estratégica más crítica.

4. **La reflexión** rememora la acción tal como ha quedado registrada a través de la observación, pero es también un elemento activo. La reflexión pretende hallar el sentido de los procesos, los problemas y las restricciones que se han manifestado en la acción estratégica. Toma en consideración la gran variedad de perspectivas que pueden darse en la situación social y permite entender las cuestiones y las circunstancias en que surgen. La reflexión se ve ayudada, habitualmente, por la discusión entre los participantes.

La investigación-acción es un proceso dinámico en el que esos cuatro momentos no deben ser entendidos como pasos estáticos, completos en sí mismos, sino como momentos en la espiral de investigación-acción constituida por la planificación, la acción, la observación y la reflexión como se puede en la Figura 1.1. Las mejoras en la comprensión se mostrarán, al comienzo, en forma de una argumentación mejor orientada hacia la práctica.

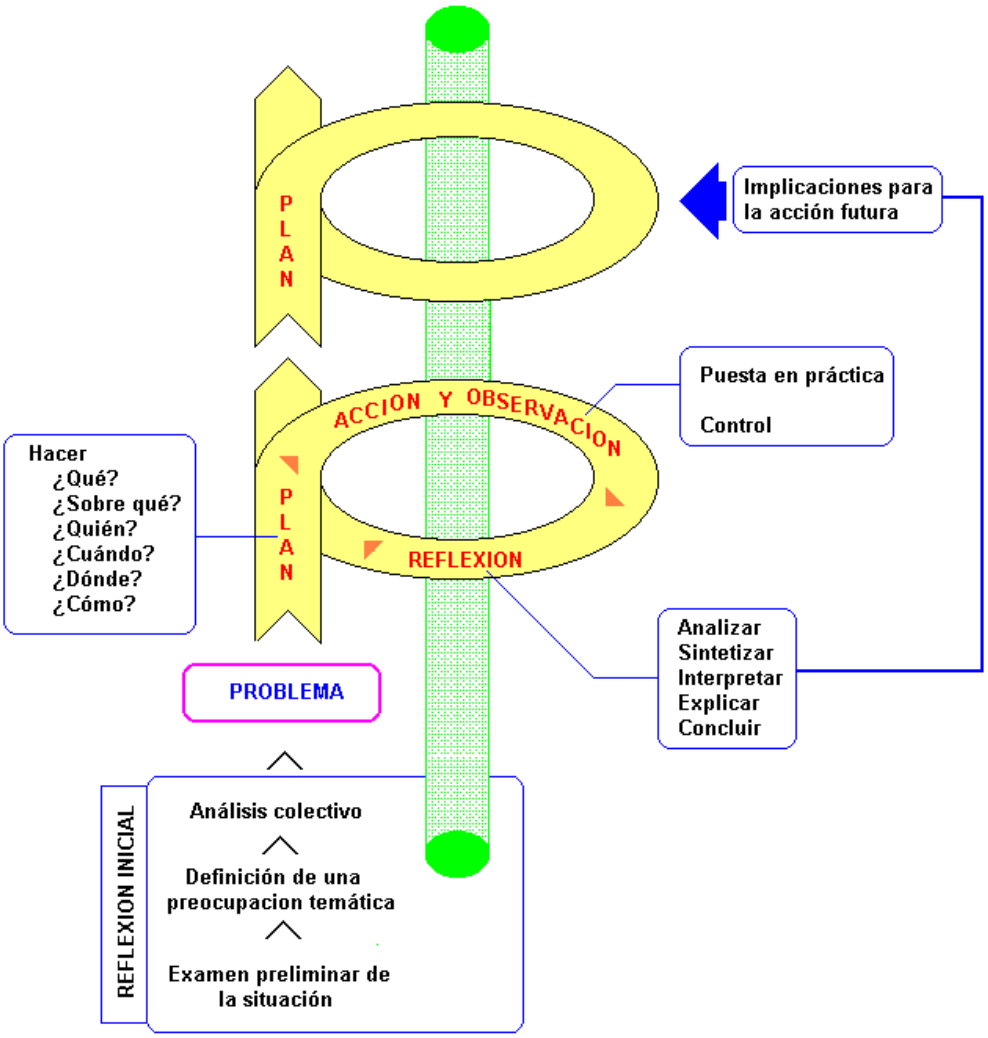

Figura 1.1: Fases de la investigación-acción

Fuente: [Barroto+02]

Varios son los ámbitos en los que el uso de este método se podrá apreciar con mayor claridad:

- **Participación en proyectos.** Algunos de los resultados del trabajo de investigación de esta tesis han sido obtenidos durante el transcurso de varios proyectos de investigación. En algunos de ellos han colaborado asociaciones y/o empresas que son participantes y, al mismo tiempo, beneficiarias de los resultados.

En este sentido, la aportación de dichos agentes ha sido de un gran valor. En concreto, cabe destacar el apoyo obtenido de la "*Asociación Vizcaína de Laringectomizados (AVL)*" que gracias a sustento se ha podido llevar a cabo esta investigación y se ha podido grabar toda la base de datos de voz esofágica. Entre los proyectos que han transcurrido en paralelo con esta investigación podemos destacar: *Oesovox* (EUROMED, Institut National de Recherche en Informatique et en Automatique, INRIA); *Darevoz* - Diagnóstico remoto por la voz a partir de medidas biométricas y otras parametrizaciones (Ministerio de Ciencia e Innovación, MICINN); Mejora de las comunicaciones telefónicas para personas con discapacidad en el habla (Proyectos de Cooperación Interuniversitaria, Túnez-España, Ministerio de Asuntos Exteriores y Cooperación, MAEC); Evaluación objetiva de patologías vocales en base a criterios acústicos y de modelado gráfico (Proyectos de Cooperación Interuniversitaria, Marruecos-España, Ministerio de Asuntos Exteriores y Cooperación, MAEC); *Dravoes* - Desarrollo del Diagnóstico, Rehabilitación y Aprendizaje de la Voz Esofágica a través de las TICs (Plan *Avanza*, Ministerio de Industria Trabajo y Comercio, MITyC); *Esofatic* (Programa *Innotek*, Departamento de Industria, Comercio y Turismo del Gobierno Vasco (GV-EJ)); Regeneración de la voz esofágica (Universidad de Deusto); *Larphone* - Evaluación objetiva de la Evolución de las Enfermedades de la Voz (Programa *Saiotek*, Departamento de Industria, Comercio y Turismo del Gobierno Vasco (GV-EJ)); *Mediproc* - Mejora de la inteligibilidad en las comunicaciones telefónicas entre laringectomizados (Programa *Saiotek*, Departamento de Industria, Comercio y Turismo del Gobierno Vasco (GV-EJ)) y *Esoimprove* - Sistema de regeneración esofágica (Programa *Saiotek*, Departamento de Industria, Comercio y Turismo del Gobierno Vasco (GV-EJ) y Universidad de Deusto).

- **Publicación en revistas internacionales y congresos internacionales.** De una forma análoga al caso anterior, las valoraciones recibidas por los

evaluadores a los que se han presentando las distintas etapas de los resultados de la investigación, han sido fundamentalmente en todas las fases de la investigación, para orientar su desarrollo.

Entre los artículos que se han publicado podemos destacar el de la revista *Computers in Biology and Medicine, Technology and Health Care* y la de *WSEAS Transactions and Systems*. En cuanto a los congresos internacionales podemos destacar: *WSEAS* conference in Signal Processing, *ISSPIT* (International Symposium on Signal Processing and Information Technology), *ICASSP* (IEEE International Conference on Acoustics, Speech and Signal Processing), *EUSIPCO* (European Signal Processing Conference), *ISSPA* (International Conference on Information Science, Signal Processing and their applications), *Biosignal* conference etc.

- **Pruebas de carácter interno.** Durante distintas etapas del desarrollo de la tesis llevaremos a cabo distintos tipos de pruebas, unas más rigurosas y otras menos, unas involucrando a más personas y otras a menos, pero en todas ellas los resultados obtenidos y el análisis de éstos permitirán obtener conclusiones sobre la validez o no de los resultados y sobre las siguientes acciones a llevar a cabo en las posteriores iteraciones del proceso cíclico.
- **Método basado en experimentos en laboratorio**

Realizar la investigación necesaria: experimentar, recopilar datos, buscar información. En el caso de esta tesis se ha recopilado una extensa base de datos de voces patológicas y laringadas. Entre las patológicas se han recogido las más graves, como son las voces esofágicas, y las menos graves, como ronquera etc.

En esta fase de experimentación, una vez recopiladas las voces esofágicas, se han desarrollado algoritmos para la mejora y la caracterización de las mismas de cara a contrastar la hipótesis planteada. Se han analizado los parámetros de la voz esofágica y se han contrastado los resultados con el programa comercial Multidimensional Voice Program (MDVP) [Deliyeski93].

ESTADO DEL ARTE

2. ESTADO DEL ARTE

"For myself, for a long time... maybe I felt inauthentic or something, I felt like my voice wasn't worth hearing, and I think everyone's voice is worth hearing. So if you've got something to say, say it from the rooftops"

Tom Hiddleston

Para conocer cómo caracterizar y mejorar las voces es necesario saber cómo se producen. En nuestro caso, es necesario ilustrar la anatomía y fisiología de la voz antes y después de la operación, es decir, antes y después de una laringectomía total. En este trabajo de investigación se estudian las diferentes estructuras del cuerpo humano que intervienen en la producción de la voz. Para ello, se presentan los conceptos generales de anatomía humana y de la voz para luego, profundizar en los elementos que integran el aparato vocal y, de modo especial, en su función de fonación. También se presenta la anatomía después una laringectomía total.

2.1 ANATOMÍA Y FISIOLOGÍA DE LA VOZ

La voz se produce gracias a la acción coordinada de casi todo nuestro cuerpo. El aparato fonador o vocal está integrado por estructuras musculares de diferentes regiones y por elementos del aparato respiratorio y del aparato digestivo. La voz se origina por una corriente aérea de los pulmones que va a ser impulsada por el

diafragma, músculos intercostales, hasta el órgano fonador. En el interior de la laringe, se encuentran las cuerdas o pliegues vocales, que trasforman el aire en sonido, mediante la vibración de sus membranas. Desde ahí va a distribuirse por el sistema resonante donde va a adquirir el timbre final de la voz [Torres08]. El aparato fonador humano lo componen tres grupos de órganos diferenciados:

- Órganos de respiración compuesto por las **cavidades infraglóticas**: pulmones, bronquios y tráquea. Es aquí donde se determinan la mayor o menor presión del aire.
- Órganos de fonación compuesto por las **cavidades glóticas**: laringe, pliegues vocales (cuerdas vocales) y resonadores nasal, bucal y faríngeo.
- Órganos de articulación compuesto por las **cavidades supraglóticas**: paladar, lengua, dientes, labios y glotis. En estos órganos el sonido producido en los pliegues vocales es amplificado y modificado.

A pesar de esta división, el aparato fonador es un todo homogéneo e inseparable, por lo cual cualquier alteración o modificación en alguna de sus partes determinará una modificación o alteración en las demás. Cualquier tensión muscular excesiva en cualquiera de ellas provocará problemas en la emisión de la voz y alteraciones a largo o corto placo en la laringe [Harries+98].

2.1.1 Cavidades Infraglóticas

El conjunto formado por los pulmones y la musculatura que suministra la energía necesaria al aire espirado ha sido denominado el fuelle del aparato fonador o vocal y componen los órganos de la respiración.

2.1.1.1 Tráquea y Pulmones

La tráquea se sitúa anterior al esófago. Se extiende entre la laringe y los bronquios principales, derecho e izquierdo, donde se bifurca. Su función es la de conducir el aire hacia los pulmones o fuera de ellos. Los pulmones son los órganos de la respiración, su función básica es la de oxigenar la sangre. Son elásticos, suaves, esponjosos y flotan en el agua. En la inspiración, la capacidad

de la cavidad torácica aumenta en las tres direcciones del espacio (por los movimientos costales y el descenso del diafragma). Al ensancharse el pulmón, se produce una reducción de la presión intraalveolar y el aire es inspirado hacia el interior. El aire entra en el pulmón como lo hace un líquido al interior de una jeringuilla al estirar del émbolo. La espiración normal o tranquila es un proceso pasivo. En la espiración activa, como durante la fonación, intervienen diversos músculos.

2.1.1.2 Diafragma. Inspiración

El diafragma es el músculo principal de la inspiración. Se sitúa como una lámina que separa la cavidad torácica de la abdominal. Tiene forma de doble cúpula y constituye el suelo de la cavidad torácica y el techo de la abdominal. Cierra la abertura inferior de la caja torácica donde se inserta. Su cara craneal es convexa y su cara caudal cóncava. Durante la inspiración, se contrae descendiendo, y durante la espiración se relaja ascendiendo.

2.1.1.3 Músculos Abdominales. Espiración Controlada

El abdomen es la porción del tronco comprendido entre el tórax y la pelvis. Durante la inspiración esta musculatura se relaja y el diafragma se contrae. En la espiración activa, los músculos del abdomen se contraen mientras que el diafragma se relaja. Esta acción coordinada constituye el denominado soporte de la voz. En la fonación, la contracción de la musculatura del abdomen provoca el aumento de la presión intra-abdominal. Este ascenso del diafragma empuja los pulmones y determina un aumento de la presión subglótica ya que los pliegues vocales, se encuentran acercados impidiendo el paso del aire. Finalmente, la presión es suficiente y el aire es espirado con fuerza produciéndose la abertura y vibración de los pliegues vocales (cuerdas vocales).

2.1.2 Cavidades Glóticas. La laringe

La laringe tiene la función de proteger las vías respiratorias y de producir los sonidos bajo la acción del aire espiratorio. Se sitúa en la parte medial y anterior

del cuello, por delante de la faringe. Interviene en la respiración, la deglución y la fonación. La laringe está formada por el hueso hioides y por los cartílagos tiroides, cricoides, aritenoides, corniculado, cuneiforme y la epiglotis y, por cuatro pares laterales, todos ellos articulados, revestidos de mucosa y movidos por músculos. La posición de estos pliegues determina la presencia de tres regiones distintas en el interior de la laringe:

- **Compartimiento superior o vestíbulo de la laringe**: es el espacio situado por encima de los pliegues vocales (cuerdas vocales).
- **Compartimento medio**: es el espacio en el que está incluido la glotis. La glotis es la porción de la laringe donde se produce la voz, e incluye los pliegues vocales y el espacio comprendido entre ellas y de los cartílagos Aritonoides (cartílago Cicrotiroideo) denominado hendidura glótica.
- **Compartimento inferior o región infraglótica**: es el espacio situado por debajo de los pliegues vocales (cuerdas vocales).

2.1.2.1 Cartílagos y Articulaciones de la Laringe.

La laringe está formada por un esqueleto de piezas cartilaginosas que se articulan entre sí. Los cartílagos de la laringe son nueve: tres impares (tiroides o nuez o bocado de Adam, cricoides y epligotis) y tres pares (aritenoides, corniculados o de Santorini y cuneiformes o de Wrisberg o de Morgagni). En muchas referencias de la literatura sólo se presentan los cartílagos principales de la laringe: Tiroides, Cricoides, Aritonoides y Epiglotis [Feneis94]. En los Aritonoides se insertan el ligamento y el músculo vocal que son el esqueleto del pliegue vocal (cuerda vocal). En estos cartílagos junto con otros se producen dos movimientos de deslizamiento. Estos movimientos determinarán la aducción (acercamiento) o abducción (separación) de los pliegues vocales.

2.1.2.2 Musculatura Intrínseca de la Laringe

En la laringe podemos distinguir una musculatura intrínseca, que determina los movimientos de las articulaciones laríngeas. Son los siguientes:

- Cricotiroideo: alarga, tensa y aduce los pliegues vocales.
- Vocal: constituye la mayor parte del pliegue vocal. Es el responsable de sus variaciones locales de tensión durante la fonación.
- Tiroaritenoideo: algunas de las fibras se extiende hasta la epiglotis formando el músculo tiroepiglótico. Es aductor de los pliegues vocales.
- Aritenoideos (transverso y oblicuo): estos músculos han sido agrupados clásicamente en la literatura bajo el nombre de aritenoideo. Es un músculo de los pliegues vocales.

2.1.2.3 Pliegues Vocales (Cuerdas Vocales)

A cada lado de la superficie interna de la laringe se encuentran dos pliegues de su mucosa superpuestos: los **pliegues vestibulares** (pliegues ventriculares, cuerdas vocales falsas, cuerdas vocales superiores o bandas ventriculares) situados cranealmente y los **pliegues vocales** (cuerdas vocales, cuerdas vocales verdaderas o cuerdas vocales inferiores) en posición caudal. El pliegue vestibular recubre el ligamento vestibular y se forma a causa de su presencia. El pliegue vocal recubre el ligamento vocal y el músculo vocal y viene determinado por la existencia de estas estructuras que forman su esqueleto. En los pliegues vocales, al paso del aire espirado, se produce un tono complejo que será modificado y amplificado en las cavidades de resonancia supraglóticas. Sin ellas el sonido producido no sería una voz laringada o sana.

2.1.2.4 Fonación. Tono, Timbre e Intensidad de la Voz

La fonación exige un cierre y una abertura continuas de los pliegues vocales (cuerdas vocales) con cambios en la longitud y la tensión. Estas variaciones requerirán fluctuaciones continuas de la salida de aire. En el habla sana o normal la regulación de la salida de este aire es básicamente voluntaria y automática. En los conferenciantes, actores y cantantes se observa, sin embargo, un control en la espiración que se ejerce por acción de los músculos abdominales. La voz es producida por la espiración del aire a través de la hendidura glótica (glotis) cerrada; los pliegues vocales son obligados a separarse y a ponerse en vibración

por la presión del aire espirado (presión subglótica) ejercida. El sonido producido en los pliegues vocales sería prácticamente inaudible si éste no se modificara y ampliara en las cavidades supraglóticas o resonadores de la voz [Laitman86]. Inmediatamente antes de la fonación (período prefonatorio) los pliegues vocales (cuerdas vocales) han de estar en contacto (aducidos) manteniendo la hendidura glótica (glotis) cerrada de modo que se interponga al paso del aire espirado. A medida que el aire intrapulmonar es expulsado se produce un aumento progresivo de la presión subglótica o infraglótica. Cuanto esta presión es superior a la de cierre de los pliegues vocales, éstos son obligados a separarse (abducirse) y el aire sale con fuerza produciéndose un descenso brusco de la presión en la hendidura glótica. Este efecto, conocido como efecto Bernoulli, junto a la elasticidad de los pliegues vocales, determina que éstos se acerquen (aduzcan) y se cierre nuevamente la hendidura glótica (glotis). Este fenómeno se va produciendo de forma rápida y repetida determinando la vibración de los pliegues vocales y, por tanto, la producción de la voz. Se denomina ciclo fonatorio o ciclo vibratorio a cada una de las fases de abertura y cierre de los pliegues vocales. El sonido producido en los pliegues vocales es un tono complejo, que consta de una frecuencia fundamental y de sus armónicos superiores. El tono aumenta cuando los ciclos de cierre y abertura de los pliegues vocales se acortan y se repiten con más frecuencia. La onda compuesta formada en la laringe pasa a través de las cavidades supraglóticas que actúan como filtros, dejando pasar sólo aquellas frecuencias que coincidan con la de las propias cavidades de resonancia. El conjunto formado por el tono fundamental más los armónicos modificados constituyen el timbre de la voz. El tono de la voz está directamente relacionado con la longitud y el grosor de los pliegues vocales de cada individuo. Las diferencias relativas entre hombres y mujeres en cuanto a la longitud (aproximadamente 18 mm. en los hombres y 10 mm. en las mujeres) y el grosor de los pliegues vocales serían los determinantes primarios de la diferencia de tono entre individuos de ambos sexos (la frecuencia fundamental en el hombre es de unos 125 Hz y en la mujer de unos 200 Hz). La intensidad o volumen de la voz dependerá principalmente de la presión del aire espirado. La

energía con la que el aire es impulsado desde los pulmones determinará una mayor o menor amplitud vibratoria de los pliegues vocales, que provocará un aumento o disminución de la intensidad del sonido producido. Al aumentar la presión del aire espirado crece la amplitud de las vibraciones, ya que los pliegues vocales se distancian y acercan con mayor agilidad.

2.1.3 Cavidades Supraglóticas

Todas las cavidades situadas por encima de los pliegues vocales (cuerdas vocales) actúan, o pueden actuar, como cajas de resonancia de la voz. Se habla de resonadores o cavidades supraglóticas. Se pueden distinguir la boca, la faringe y las fosas nasales. Dichas cavidades funcionan a modo de filtro y resonador, atenuando y amplificando determinadas frecuencias de la onda generada por el aire que se expulsa de los pulmones y que atraviesa la cavidad laríngea. Según la forma que adopten los órganos en las cavidades supraglóticas y la zona que esté implicada obtendremos un sonido u otro.

2.1.3.1 La Faringe

La faringe es la parte del tubo digestivo situada por detrás de la cavidad nasal, la bucal y la laringe. Conecta la nariz y la boca con la laringe y el esófago respectivamente, y por ella pasan tanto el aire como los alimentos, por lo que forma parte del aparato digestivo así como del respiratorio. La faringe puede actuar como resonador de la voz. En función del tamaño de esta cavidad el aire espirado resonará en ella con mayor o menor intensidad.

2.1.3.2 Velo Paladar

El paladar constituye el suelo de la cavidad nasal y el techo de la boca. Está formado por dos porciones, el paladar duro u óseo y el velo del paladar o paladar blando. En los sonidos nasales el velo del paladar se halla relajado y el aire pasa hacia la cavidad nasal, donde resuena; en los sonidos orales el velo del paladar está elevado y tensado cerrando el paso hacia las fosas nasales, con lo cual el aire resuena únicamente en la boca.

2.1.3.3 La Boca

La boca es el principal resonador de la voz. Puede adaptar su forma y volumen al sonido emitido en los pliegues vocales (cuerdas vocales) por medio de los cambios en la posición de la lengua, los labios, el velo del paladar y la mandíbula. El sonido generado en los pliegues vocales es impartido al ambiente por medio de la boca. La intensidad o volumen final del sonido será directamente proporcional al área de abertura de ésta. Asimismo, la abertura de la boca influirá en el timbre de la voz [Boone97].

La posición de los labios influirá también en la forma de la cavidad de resonancia, y, por tanto, en el timbre de la voz. Si abrimos la boca horizontalmente, la voz tendrá un timbre o color más claro que si colocamos los labios de forma circular, en cuya situación el timbre será más oscuro.

2.1.3.4 Cavidad Nasal

La cavidad nasal puede estar separada total o parcialmente de la cavidad bucal por el velo del paladar, produciéndose los sonidos orales o nasales del habla. No se debe confundir nariz con cavidad nasal. La nariz es la porción externa que proyecta la cavidad nasal hacia delante. La cavidad nasal está formada por las fosas nasales derecha e izquierda separadas por el tabique nasal. Para que la cavidad nasal actúe como una cavidad de resonancia es necesario que el velo del paladar esté relajado y el aire espirado salga por esta región.

2.1.4 La Voz tras una Laringectomía

El cáncer de laringe es aquél en el que la laringe queda afectada o dañada por las células cancerosas. El tratamiento puede incluir una laringectomía parcial o total, es decir, en el caso de la laringectomía total la laringe se extirpa y, por lo tanto, no existe, con lo que ya no están tampoco las cuerdas vocales, ni la epiglotis, ni los cartílagos que la rodeaban. Los avances en tratamientos de cirugía, la radioterapia y la quimioterapia a menudo pueden ahora salvar la laringe o parte de ella (laringectomía parcial). Mantener la laringe mantiene la voz, incluso si su

calidad se modifica. Pero para las personas con cánceres avanzados, la eliminación de la laringe puede ser la mejor opción para preservar la vida. Si finalmente se realiza la operación, hay que aprender a hablar utilizando diferentes técnicas.

2.1.4.1 La Voz Esofágica

En la voz esofágica se toma el aire del esófago y se deja escapar por la faringe. Al realizar esta acción, la parte superior del esófago faríngeo vibra y produce el sonido. La voz producida es algo así como un eructo, pero algo diferente ya que en este caso el aire no proviene del estómago. Hablar con el esófago es difícil y conlleva un tiempo aproximado aprendizaje de hasta seis meses. La principal consecuencia de este abocamiento de la tráquea al cuello es que el aire de los pulmones no puede ya circular en la boca ni en la nariz y no hay forma a primera vista de soplar, ni hablar, ni silbar, ni sonarse. La posibilidad de hablar en voz alta la han perdido por dos razones: el vibrador ha sido extirpado (no existe laringe ni cuerdas vocales para sonorizar el aire) y el soplo ha sido desempalmado (el aire de los pulmones ha sido desviado de su trayecto normal). La producción de voz esofágica no requiere ningún tipo de operación adicional o prótesis especial y, es además, relativamente fácil de aprender. Sin embargo, como ya se ha mencionado anteriormente, la voz es parecida a un eructo y se limita a la producción de segmentos y frases cortas de habla, debido a la continua necesidad de tener que volver a llenar de aire la parte alta del esófago. Además tiene una componente de ruido, que resulta bastante molesta para el interlocutor, sobre la cual hay diversos trabajos de investigación [Javkin+97] [Sakakibara+02].

2.1.4.2 La Voz Traqueosofágica

En el habla traqueosofágica el cirujano realiza una punción en la que se inserta una prótesis de voz. El aire es exhalado y pasa desde el pulmón al esófago a través de la tráquea mediante la prótesis implantada. De esta manera, la faringe inferior vibra produciendo la voz. Para desviar el aire espirado a la prótesis ésta

debe ser ocluida. El sonido que podrá ser modulado con el movimiento de labios, lengua, boca...etc.

2.1.4.3 Laringe Artificial

Otra forma de comunicarse es mediante una Laringe Artificial. La vibración se produce por un vibrador eléctrico externo también llamado electrolaringe. Hay dos tipos laringes: de tipo cuello o intra-oral. El tipo de cuello se coloca sobre la piel en el lado del cuello, debajo de su barbilla, o en su mejilla. Se trata de un dispositivo que contiene un diafragma vibrador. Esto aparato produce una vibración en la garganta que duplica a la de las cuerdas vocales. La voz producida de esta forma tiene un sonido metálico, pero puede ser fácilmente entendible, incluso por teléfono. Con este método existe la posibilidad de hablar frases largas y entendibles. No necesita ningún cuidado especial, simplemente ha de colocarse en el cuello y encenderlo. Además, puede ser utilizado por casi todos los pacientes. La pega es que el tono de voz obtenido de esta forma es muy mecánico y no suena natural debido a la falta de variación la frecuencia fundamental o Pitch (Jitter) y a la variación de amplitud (shimmer).

2.2 TÉCNICAS DE PROCESADO DE SEÑAL PARA LA MEJORA DE LA VOZ

Las técnicas de procesado de señal de la voz son aplicadas a un modelado de la voz. Como ya se ha comentado en la sección anterior, la voz se produce insuflando una columna de aire que proviene de los pulmones y excita el tracto vocal comportándose ésta como una cavidad resonante. Desde el punto de vista de procesado de señal, los sonidos emitidos pueden clasificarse de dos formas:

2.2.1 Sonoros o tonales (voiced)

Cuando se emiten este tipo de sonidos se produce una oscilación relajada de las cuerdas o pliegues vocales. Tecnológicamente hablando puede pensarse que el tracto vocal es excitado por un tren de pulsos de aire cuasi-periódicos. Estas señales se caracterizan por tener una alta energía y están comprendidas entre un

rango de frecuencias entre 300 Hz y 4000 Hz. Los ejemplos más significativos de estas señales son las vocales y algunas consonantes.

2.2.2 Sordas o no tonales (unvoiced)

Cuando se producen este tipo sonidos las cuerdas permanecen abiertas y el aire proveniente de los pulmones llega con la suficiente velocidad que produce una turbulencia. Estas señales se caracterizan por tener una baja energía. Las frecuencias de estas señales están repartidas en todo el espectro de frecuencias uniformemente y son de forma aleatoria muy parecidas al ruido blanco. Las consonantes fricativas son ejemplos de estas señales.

2.2.3 Modelo del Tracto Vocal

Los diferentes bloques que toman parte de la producción del habla son modelados de la siguiente manera: las cavidades subglóticas se modelan como la fuente de excitación del sistema, la cavidad laríngea y los pliegues vocales se representan por medio del filtro de tracto glótico y, por último, las cavidades supraglóticas son modeladas por el filtro de tracto vocal.

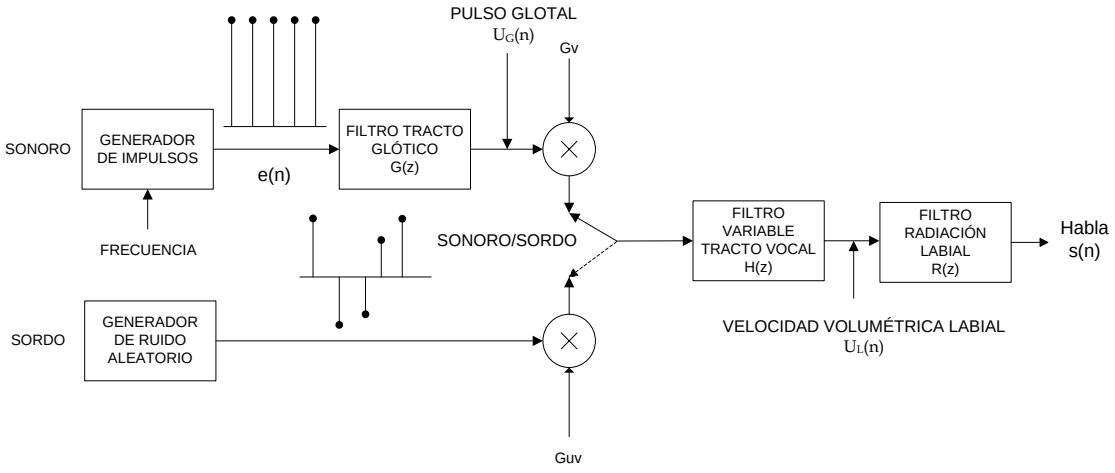

Figura 2.1: Modelo de tracto vocal para la producción del habla

Al final de este sistema lineal se encuentran las fosas nasales y los labios, representados por el filtro de radiación labial. A continuación se describen las señales y los filtros que intervienen en el modelo del tracto vocal:

- e(n): Modelo de señal de excitación glótica.
- G(n): Respuesta impulsional del filtro de tracto glótico.
- U_G(n): Pulso glótico o señal de velocidad volumétrica glótica.
- H(z): Función de transferencia del filtro de tracto vocal.
- U_L(n): Señal de velocidad volumétrica labial.
- R(z): Función de transferencia del filtro de radiación labial.
- s(n): Señal de presión acústica.

2.2.3.1 Modelo de tracto glótico

La entrada del filtro de tracto glótico será un tren de impulsos de frecuencia igual a la frecuencia fundamental de la voz (F_0), con una respuesta impulsional del filtro que corresponde con el pulso glótico. En el caso de sonidos sordos, este filtro no se utilizará, usando como excitación del filtro de tracto vocal una señal de ruido blanco no correlado.

Existen varios modelos de la respuesta de la glotis por un tren de pulsos: modelo exponencial [Krishnamurthy92], modelo de Rosenberg [Rosenberg71] [Qiang06] y modelo Liljencrants-Fant (LF) [Fant+85].

Si fijamos nuestra atención en el modelo exponencial, la función de transferencia del filtro es:

$$G(z) = \frac{-ae\ \ln(a)z^{-1}}{(1-az^{-1})} \tag{2.1}$$

donde a es el módulo de polos de la función de transferencia.

El modelo propuesto por [Fant+85], el modelo Liljencrants-Fant (LF) es una representación del flujo glótico y su derivada. En el modelo de fuente del mecanismo de producción del habla, el flujo glótico sirve como excitación para el

filtro de tracto vocal. Ha sido un modelo utilizado en estudios analíticos. El modelo de derivada del pulso glótico viene dado por la suma de las siguientes ecuaciones:

$$E_1(t) = E_0 e^{\alpha t} \sin(\omega_g t) \; : \; 0 < t < t_e \tag{2.2}$$

$$E_2(t) = -\omega_g K U_0 \sin\left(\omega_g (t - t_p)\right) \; : \; t_e < t < t_c \tag{2.3}$$

donde los cuatro parámetros propuestos son la amplitud (E_0), la frecuencia (ω_g), la constante de crecimiento exponencial de la sinusoide (α) y la constante de recuperación exponencial. Además, el intervalo de tiempo $[0, t_e)$ se refiere a la fase abierta de la glotis, es decir, cuando la glotis se abre y se cierra. El intervalo $[t_e, t_c)$ corresponde a la fase de retorno del modelo LF donde la glotis se está cerrando con su máximo en t_p. A partir de este intervalo la glotis se considera que está cerrada $[t_e, t_0)$.

Rosenberg propuso un filtrado inverso para la extracción del modelo de onda glótica de la señal del habla. Este modelo está definido por dos polinomios:

$$g(t) = \begin{cases} t^2(t_e - t), 0 < t < t_e = t_c \\ 0 \quad , \quad t_c < t < t_0 \end{cases} \tag{2.4}$$

donde los parámetros de los intervalos son los mismo anteriormente mencionados en el modelo LF.

2.2.3.2 Modelo de tracto vocal

El tracto vocal se puede modelar como la concatenación de tubos acústicos de distintos diámetros (con o sin pérdidas) [Bäckström04]. Esto deriva en un **modelo lineal no estacionario** (ya que las secciones de los tubos van cambiando de acuerdo al fonema que se está emitiendo) [Rabiner+98]. El tracto vocal modelado se manifiesta como un filtro variable en el tiempo cuyos parámetros varían en el tiempo en función de la acción consciente que se realiza al pronunciar una palabra o un fonema [Abramovich+07]. El filtro variable en el tiempo tiene dos

posibles señales de entrada que dependerán del tipo de señal, sonora o no sonora. Para señales sonoras la excitación será un tren de impulsos de frecuencia controlada, mientras que para las señales no sonoras la excitación será ruido aleatorio. El espectro de frecuencias de la señal vocal puede obtenerse a partir del producto del espectro de la excitación por la repuesta en frecuencia del filtro. El tracto vocal manifiesta un número muy grande de resonancias, sin embargo se consideran solo las tres o cuatro primeras que toman el nombre de formantes y cubren un rango de frecuencias entre 100 Hz y 3500 Hz. Los formantes están localizados en diferentes posiciones dependiendo del sonido producido. Sin tener en cuenta el retraso introducido ni las pérdidas, la función de transferencia resultante del tracto vocal es un filtro todo-polos, con N polos determinados por los coeficientes a_k donde el índice podrá tener los siguientes valores $k = 1, 2, 3 \cdots N$:

$$H(z) = \frac{1}{1+\sum_{k=1}^{p} a_k z^{-k}} \tag{2.5}$$

donde *p* es el orden del filtro todo-polos.

Esta función de transferencia proviene del clásico modelo de autoregresión lineal del habla (Autoregressive (AR) Model of Speech, [Proakis+07]). Esta función de transferencia proviene del modelo de voz en el que se parte de la idea de que la voz se predice como la combinación lineal de las p muestras anteriores más una señal de error:

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + \omega(n) \tag{2.6}$$

donde a_k se denominan los coefcientes LPC (Linear Predictive Coding) y p es el orden del filtro. Los coeficientes de predicción lineal son determinados mediante la información previa de la señal de voz y el orden p son las observaciones previas utilizadas para la estimación. La manera más común de la estimación de estos parámetros es mediante el método de mínimos cuadrados o método de autocorrelación de Yule-Walker [Walker31] [Yule27], junto con el método de

Levinson [Levinson47]. Los coeficientes LPC se consideran constantes pero existe variantes de este modelo haciendo que los coeficientes varíen en el tiempo o que en vez de tener un sistema todo-polos, tener una función de transferencia con polos y ceros [Lewis+08]. Algunos ejemplos de estos modelos son: Time-varying models (TVAR), Autoregressive model with exogenous input (ARX), Vector models (VAR), Autoregressive moving average models (ARMA) etc.

2.2.3.3 Modelo de radiación labial

La onda de la señal de voz $s(n)$ está relacionada con la onda de velocidad volumétrica $u_L(n)$ presente en los labios a través de una impedancia de radiación R(z) [Wong+79], que se considera invariante de acuerdo al sonido producido.

Para frecuencias por debajo de los 4000 Hz, la señal de presión sonora, a una distancia l de los labios, es proporcional a la derivada temporal de la velocidad volumétrica en los labios con un tiempo de retraso de l/c_0. Excluyendo la constante de proporcionalidad y el tiempo de retraso, a bajas frecuencias se puede aproximar la impedancia de radiación R(z) por un filtro pasa-alto.

$$R(z) = 1 - \alpha z^{-1} \tag{2.7}$$

2.2.4 Transformada Wavelet

Es bien conocido de la teoría de Fourier [Fourier22] que una señal puede ser expresada como la suma, posiblemente infinita, de serie de senos y cosenos. Esta suma también se conoce como un desarrollo de Fourier. La gran desventaja de un desarrollo de Fourier, sin embargo, es que tiene resolución frecuencial y sin resolución de temporal. Esto significa que aunque podría ser capaz de determinar todas las frecuencias presentes en una señal, no se sabe cuándo están presentes. Para superar este problema, en las últimas décadas, han sido desarrolladas varias soluciones, que son más o menos capaces de representar una señal en el dominio del tiempo y de la frecuencia al mismo tiempo.

La idea que subyace de estas representaciones de tiempo-frecuencia es tomar fragmentos de la señal de interés y luego analizar dichos fragmentos por separado. Está claro que este tipo de análisis en una señal aportará más información y tendrá una mayor resolución tiempo-frecuencia. El problema es que no es posible obtener una resolución absoluta de tiempo-frecuencia debido al principio de incertidumbre de Heisenberg que, en términos de procesamiento de señales, indica que es imposible saber la frecuencia exacta y el momento exacto de aparición de una frecuencia concreta en una señal. En otras palabras, una señal no puede simplemente ser representado como un punto en el espacio de tiempo-frecuencia.

La transformada wavelet o análisis wavelet es una de las soluciones para superar las deficiencias de la transformada de Fourier. En el análisis wavelet el uso de una ventana modulada totalmente escalable resuelve el problema del fragmento seleccionado. La ventana se desplaza a lo largo de la señal y para cada posición del espectro se calcula. Entonces este proceso se repite muchas veces con diferentes tamaños de ventana para cada nuevo ciclo. Al final, el resultado será una colección de representaciones tiempo-frecuencia de la señal, todos ellos con diferentes resoluciones. Debido a esta colección de representaciones se puede hablar de un análisis multiresolución. En el caso de las wavelet, normalmente, no se habla de representaciones tiempo-frecuencia, sino de las representaciones a tiempo-escala, donde la escala es el inverso de la frecuencia. Esto es debido a que la frecuencia es un término que se reserva para la transformada de Fourier [Ortolan+03].

2.2.4.1 Transformada Wavelet Continua

La transformada wavelet continua, *Continuous Wavelet Transform (CWT)*, se define matemáticamente de la siguiente manera [Akansu 92] [Akansu+10]:

$$\gamma(s,\tau) = \int f(t)\, \psi^*_{s,\tau}(t) dt \qquad (2.8)$$

donde $\psi^*_{s,\tau}(t)$ son la base de wavelets madre conjugadas, la variable *s* es la escala, la nueva dimensión y la variable τ es la traslación. La transformada wavelet inversa es [Allingham+98]:

$$f(t) = \iint \gamma(s,\tau)\, \psi^*_{s,\tau}(t) d\tau ds \tag{2.9}$$

Las wavelets se generan a partir de una única wavelet base $\psi(t)$, llamada wavelet madre, mediante el escalado y la traslación:

$$\psi^*_{s,\tau}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-\tau}{s}\right) \tag{2.10}$$

Todas las wavelet son generadas a partir de la wavelet madre y el factor $\frac{1}{\sqrt{s}}$ es simplemente un factor de normalización.

Para que una función sea considera una wavelet madre debe cumplir algunas condiciones: la condición de admisibilidad y regularidad. Para que la transformada wavelet sea reversible sin pérdida de información debe cumplirse la condición de admisibilidad que se muestra a continuación [Sheng96a]:

$$\int \frac{|\psi(\omega)|^2}{|\omega|} d\omega < \infty \tag{2.11}$$

donde $\psi(\omega)$ es la transformada de Fourier de la función wavelet madre $\psi(t)$. La ecuación anterior implica que la transformada de Fourier de la función $\psi(t)$ se hace cero si la frecuencia $\omega = 0$:

$$|\psi(\omega)|^2 = 0 \quad si\ \omega = 0 \tag{2.12}$$

Esto nos dice de las wavelets que tiene un "comportamiento" parecido a los filtros paso-banda. Además se puede deducir que:

$$\int f(t) dt = 0 \tag{2.13}$$

Es decir, para que la función sea una onda, dicha función debe ser oscilatoria en todo el dominio de la dicha función.

La regularidad es un concepto un tanto más complejo y para explicarlo introduciremos el concepto de "*vanishing moments*". Si se realiza la expansión de la serie de Taylor de la transformada wavelet para $\tau = 0$, aparecen los en todos sus términos la siguiente componente [Chui92]:

$$M_p = \int t^p \ \psi(t) dt \tag{2.14}$$

donde a los M_p se denominan los momentos de la wavelet madre. De la propiedad de admisibilidad, sabemos que $M_p = 0$ si $p = 0$. Si ahora se consigue que los siguientes M_n momentos sean cero, los coeficientes de la transformada wavelet, $\gamma(s, \tau)$, decaerán con el término del polinomio de orden s^{n+2}. Esto se conoce como los *vanishing moments* u orden de aproximación. Si una wavelet tiene *N vanishing moments* entonces el orden de aproximación será *N* [Calderbank+96]. De estas dos propiedades le viene el nombre de wavelet: de la admisibilidad proviene la propiedad de *onda* y de regularidad onda *pequeña*.

Existen familias de wavelets y cada una de ellas posee unas características diferentes a las otras. Entre las familias más conocidas podemos destacar: Morlet, Mexican hat, Meyer, Haar, Daubechies, Symlets, Coiflets y Splines biorthogonal wavelets (bior Nr.Nd donde Nr y Nd son los órdenes asociados a los filtros de reconstrucción y descomposición). Aquí podemos ver algunas de las wavelets madre más populares para unos órdenes concretos [Daubechies92] [Chui92] [Lió03]:

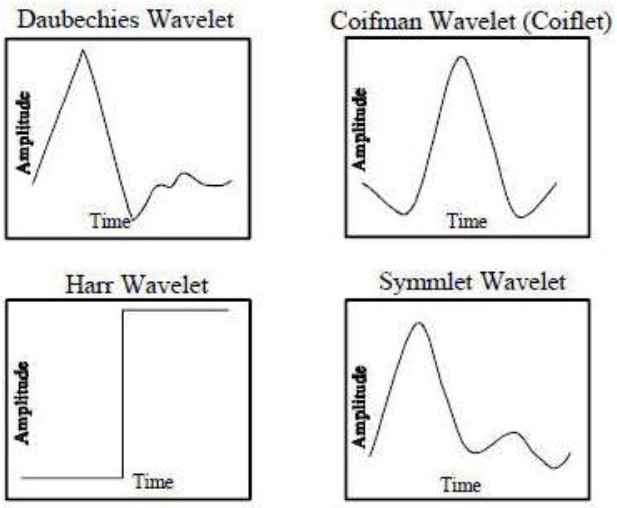

Figura 2.2: Algunos ejemplos de wavelet madre

Fuente: http://www.rfcafe.com/references/electrical/ew-radar-handbook/transforms-wavelets.htm

En la transformada wavelet hay información relativa al tiempo y a la frecuencia (escala). Debido a ello, se representa esta información en tiempo-frecuencia o tiempo-escala. Esta es precisamente, una de las ventajas de la transformada wavelet, como ya se ha comentado anteriormente, que la resolución en el plano tiempo-frecuencia o tiempo-escala es mejor que la que se utilizaba anteriormente, la transformada de Fourier (o también Short Time Fourier Transform, STFT) [Jacobs93].

Figura 2.3: Representación de tiempo-frecuencia del STFT y WT

Fuente: Matlab Wavelet Toolbox

Las áreas de las celdas tanto para la STFT como para la transformada wavelet son las mismas y están determinadas en el principio de incertidumbre de *Heisenberg*.

2.2.4.2 Transformada Wavelet Discreta

Una vez conocida las características de la transformada wavelet continua (CWT), se puede ver que es poco práctica para utilizarla en aplicaciones reales. La transformada wavelet continua descrita tiene tres propiedades que la hacen difícil de usar [Daubechies92] [Calderbank+96] [Mallat99]:

- La redundancia de la transformada wavelet continua (CWT). Como ya hemos mencionado anteriormente, la transformada wavelet se calcula para una translación continua y una función escalable continua sobre toda la señal, calculando la correlación entre ambas. Está claro que la función escalada dista mucho de ser una base ortogonal (como los senos y cosenos de Fourier) y, por lo tanto, los coeficientes wavelets son altamente redundantes.

- Tenemos infinitas wavelets en la transformada wavelet continua (CWT). Hay que reducir este número a un volumen manejable de wavelets si se desea que sea operativa.

- Para la mayoría de las aplicaciones, la transformada wavelet continua (CWT) no tiene solución analítica y sólo puede ser calculada numéricamente.

Para reducir esta redundancia nace la transformada wavelet discreta (DWT). Para construir dicha transformada wavelet se realiza la discretización de los parámetros de tal manera que la información de la transformada discreta sea manejable y fácilmente computable. Concretamente se toma esta discretización:

$$\begin{cases} s = s_0^j \\ \tau = k\tau_0 s_0^j \end{cases} \tag{2.15}$$

donde los índices *j,k* son enteros. De esta forma la wavelet quedan representadas de la siguiente manera[Mallat89] [Burrus+98]:

$$\psi_{j,k}(t) = \frac{1}{\sqrt{s_0^j}}\psi\left(\frac{t - k\tau_0 s_0^j}{s_0^j}\right) \tag{2.16}$$

A esta función se le llama la wavelet discreta (o semidiscreta) pero es una discretización de la función continua. Este efecto hace que el espacio tiempo-escala esté muestreado en intervalos discretos. Si $s_0 > 1$ es un paso de dilatación fijo. Se suele elegir una potencia de dos para la escala $s_0 = 2$ y $\tau_0 = 1$ para el factor de traslación. Esto hace que la representación tiempo-escala sea *diádica* (ver Figura 2.4).

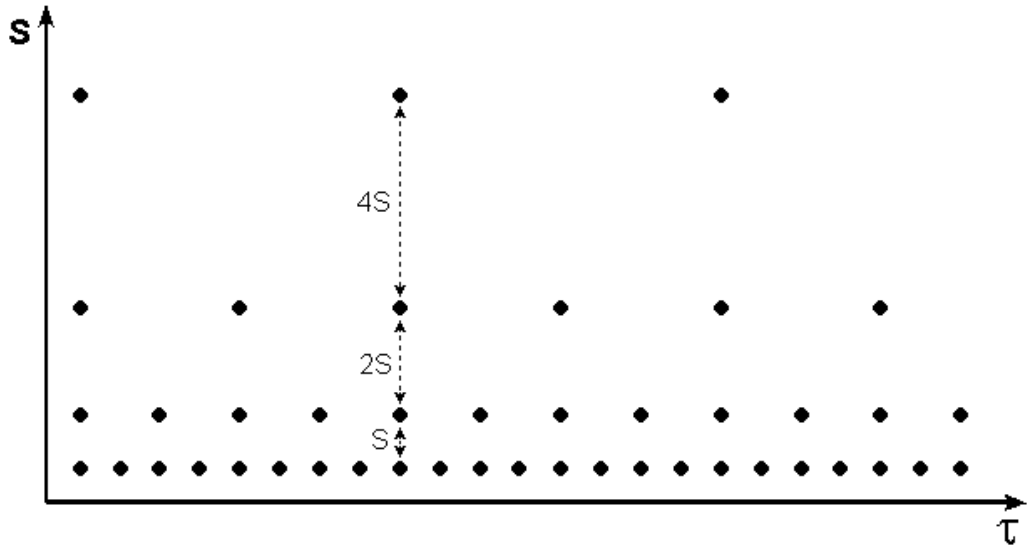

Figura 2.4: Representación gráfica de tiempo-escala en red diádica

Fuente: http://www.polyvalens.com

Cuando se aplica la transformada wavelet discreta (DWT), se descompone en series de coeficientes de wavelets. Estos coeficientes sirven para la reconstrucción de la señal.

Es bien conocido que para la reconstrucción estable es condición necesaria y suficiente, que el cuadrado de la energía de la los coeficientes wavelets tiene que estar entre dos límites positivos [Chui92] [Daubechies92] [Sheng96b]. Esto es:

$$A\|f\|^2 \leq \sum_{j,k}\left|\langle f, \psi_{j,k}\rangle\right|^2 \leq B\|f\|^2 \tag{2.17}$$

donde los valores A y B son $A > 0$ A > 0 y $B > \infty$ e independientes de $f(t)$. La energía de $f(t)$ es $\|f\|^2$. La familia de base de funciones $\psi_{j,k}(t)$ con $j, k \in \mathbb{Z}$ es referida como *ventana* con los límites A y B. Cuando $A = B$, la transformada wavelet discreta (DWT) se comporta exactamente como una base de funciones ortonormales. Cuando $A \neq B$ todavía es posible la exacta reconstrucción de la señal original a costa de una *ventana dual*. Es decir, con la *ventana dual*, en la transformada wavelet discreta (DWT), la descomposición wavelet es diferente a la reconstrucción wavelet.

Para hacer que la transformada wavelet sea menos redundante, el último paso es hacer que la transformada wavelet discreta ortonormal. Esto sólo es posible con las wavelets discretas. Para que las wavelets discretas sean ortonormales con respecto a las versiones wavelets escaladas y trasladadas es necesario escoger de especial forma las wavelets madre. Es decir, se debe cumplir lo siguiente [Cohen+92] [Daubechies88]:

$$\int \psi_{j,k} \psi^*_{m,n}\, dt = \begin{cases} 1, & j = m\,;\, k = n \\ 0, & j \neq m\,;\, k \neq n \end{cases} \tag{2.18}$$

Esto supone que al igual que en la transformada de Fourier, una señal arbitraria puede ser representada por la suma de una base de funciones wavelets ortogonal ponderada por los coeficientes de la transformada wavelet [Strang96]:

$$f(t) = \sum_{j,k} \gamma(j,k)\, \psi_{j,k}(t) \tag{2.19}$$

donde los coeficientes wavelets vienen dados por $\gamma(j,k)$ y la ecuación anterior es la inversa de la transformada discreta wavelet. Si la familia wavelets, $\psi_{j,k}(t)$, son ortonormales o biortogonales, la transformada no será redundante. Si por el contrario, la wavelet es una *ventana*, la transformada habrá redundancia [Daubechies88] [Vaidyanathan+01].

A pesar de se ha reducido el número de wavelets en la transformada, y con ello la redundancia, aún se necesitan infinitas wavelets dilatadas y escaladas para el

cálculo de la transformada. Además, persiste el problema de la resolución analítica de la transformada.

Si analizamos el problema del número finito de wavelets con respecto a la traslación se deduce que el número de wavelets viene dado por la longitud de la señal original. Es decir, tiene un límite superior finito. Esto reduce el problema a la dilatación. Esto nos conduce a introducir el concepto de la *función de escalamiento* (*scaling function*) [Mallat89].

Como ya se ha mencionado en anteriormente, la propiedad wavelet mostrada de la ecuación 2.12 nos dice que la función wavelet se comporta como un filtro paso-banda. Si a esta propiedad le añadimos que cuando se produce una compresión de una función en el tiempo tanto su ancho de banda como sus componentes frecuenciales aumentan en el mismo orden, entonces podemos ver las diferentes wavelets dilatadas como un banco de filtros paso-banda (Figura 2.5).

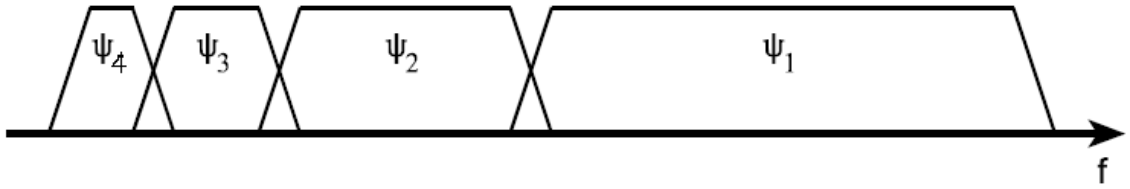

Figura 2.5: Anchos de banda con respecto a la dilatación de la wavelet madre.

Fuente: http://www.polyvalens.com

En el caso de la Figura 2.5, como la traslación depende de la escala, se comienza el barrido de escalas menores (frecuencias mayores) a escalas mayores (frecuencias menores). Esto es coherente con que en la traslación también se produce un barrido de una menor dilatación a una mayor dilatación, ya que como hemos visto anteriormente una depende de la otra, ecuación 2.15, dando lugar a menores anchos de banda y menores componentes frecuenciales a medida que tenemos wavelets dilatadas. Es decir, que cada vez que se dilate la wavelet por un factor 2, el ancho de banda se reducirá a la mitad. A la hora de diseñar una wavelet se debe tener en cuenta este hecho, es decir, que los filtros tengan un ancho de banda que se superpongan entre sí y que cubran el mayor

ancho de banda posible. Los filtros utilizados son filtros de espejo en cuadratura [Vaidyanathan87] [Vetterli87].

Esto no quiere decir que se vaya a cubrir todo el espectro con las versiones dilatadas de las wavelets. Es decir, que por cada wavelet dilatada únicamente se cubre la mitad del espectro restante y, por lo tanto, se siguen necesitando infinitas wavelets. Para solucionar este hecho Mallat introdujo la *función de escalamiento* [Mallat89]. La función de escalamiento tiene un ancho de banda en modo filtro paso-bajo que cubre el "hueco" dejado por las wavelets dilatadas. Además, debido a que el ancho de banda de la *función de escalamiento* es finito y que cualquier función se puede expresar por medio de la base de funciones wavelet (ecuación 2.20), el número de wavelet es finito con lo que el problema está resuelto [Mallat99].

$$\varphi(t) = \sum_{j,k} \gamma(j,k)\, \psi_{j,k}(t) \qquad (2.20)$$

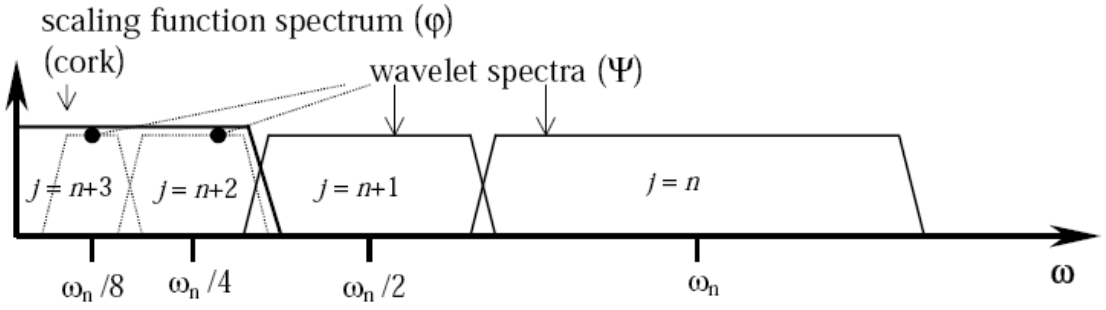

Figura 2.6: Reemplazo de infinitas wavelet por la función de escalamiento.

Fuente: http://www.polyvalens.com

Si consideramos la transformada wavelet como un banco de filtros, entonces podemos considerar la transformada de una señal como el paso de dicha señal por un banco de filtros. Las salidas de los diferentes filtros son los coeficientes de las transformadas wavelets y la transformada de la función de escalamiento. Al análisis de una señal por medio de un banco de filtros se le denomina codificación sub-banda [Vetterli+95]. De esta manera, cada vez que se pasen los

filtros a la señal se dividirá el espectro. Se puede concluir que realizar una transformada wavelet es como realizar una codificación sub-banda con filtros de espejo en cuadratura [Mallat89].

La transformada discreta wavelet analiza la señal descomponiéndola en dos: aproximación y detalle, todo ello para un nivel dado considerando diferentes bandas de frecuencia con distintas resoluciones para cada nivel. Con este propósito se emplean los dos conjuntos de funciones mencionados: funciones de escalamiento y funciones wavelets. Las funciones de escalamiento se asocian a un filtro paso-bajo $[h(n)]$ y las funciones wavelets a filtro paso-alto $[g(n)]$. Las descomposición de la señal se obtiene mediante filtrados iterativos paso-bajo y paso-alto y después se produce un sub-muestreo de la señal eliminando redundancia y de acuerdo a la regla de Nyquist-Shannon [Shannon49] (Figura 2.7). De esta manera se descompone la señal de un nivel a otro y es algoritmo utilizado para realizar la transformada discreta wavelet (DWT) que matemáticamente se expresa [Mallat99] según la ecuación (2.21).

$$\begin{cases} y_{high}(k) = \sum_n s(n) \cdot g(2k-n) \\ y_{low}(k) = \sum_n s(n) \cdot h(2k-n) \end{cases} \tag{2.21}$$

En la ecuación (2.21). $y_{high}(k)$ es la salida del filtro paso-alto e $y_{low}(k)$ la del paso-bajo tras realizar un sub-muestreo por 2. Esta operación reduce a la mitad la resolución en el tiempo, como consecuencia de la reducción a la mitad del número de muestras originales posee señal. Sin embargo, este mismo proceso duplica la resolución en frecuencia ya que ahora la banda de frecuencia de la señal abarca solamente la mitad de la banda de frecuencias anteriores [Rioul92] [Mallat99] [Strang89].

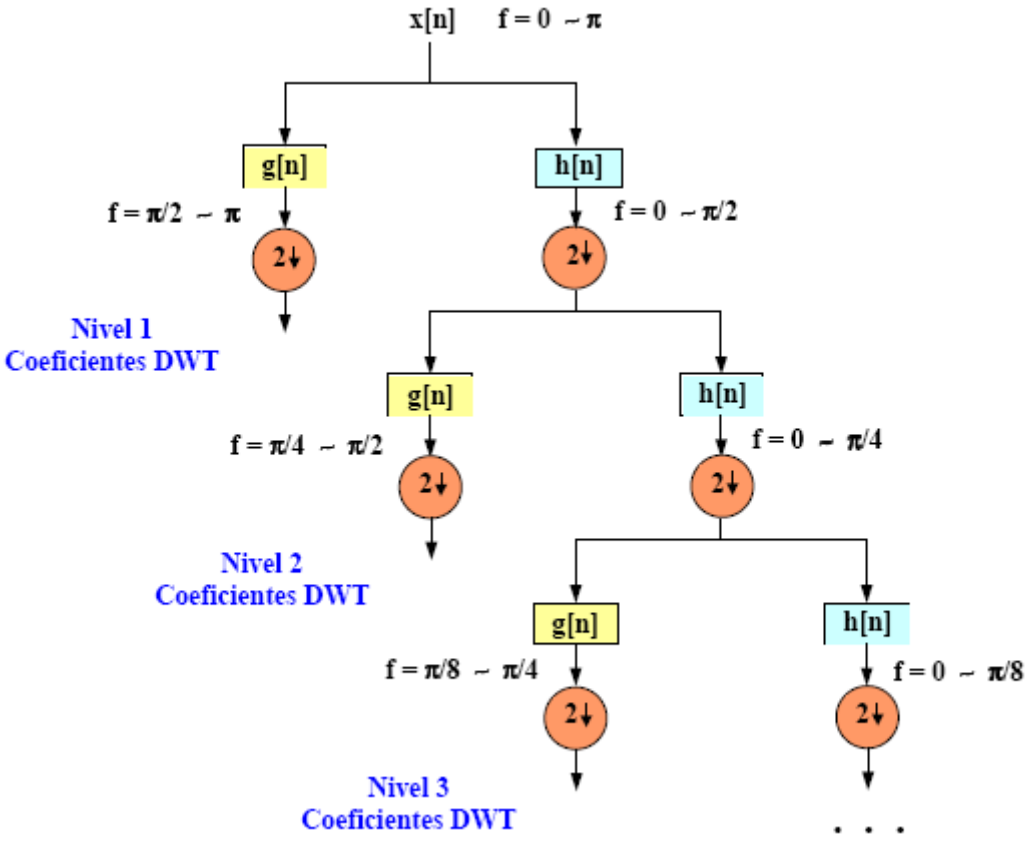

Figura 2.7: Algoritmo de codificación sub-bandas

Fuente: http://users.rowan.edu/~polikar/WAVELETS/WTpart4.html

2.2.5 Filtros de Kalman

En 1960, R.E. Kalman publicó su famoso trabajo describiendo un método recursivo para el problema de filtrado lineal de datos digitales [Kalman60]. Desde entonces, los filtros de Kalman han sido objeto de multitud de investigaciones y de numerosas aplicaciones. Una de ellas es la mejora del ruido de la señal de voz [Gannot98] [Grancharov+06] [Yang+02].

El filtro de Kalman es un conjunto de ecuaciones matemáticas que proporcionan una solución computacional eficiente (recursiva) del método de mínimos cuadrados. El filtro es muy potente en varios aspectos: es compatible con las estimaciones de estados pasados, presentes y futuros, y puede hacerlo incluso cuando la naturaleza del sistema modelado no es precisa o es desconocida. Por lo tanto, podríamos decir que el filtro de Kalman es un estimador lineal que

resuelve el problema de estimación de estados en sistemas dinámicos que están perturbados por ruido blanco.

Desde el punto de vista estadístico, este estimador nos proporciona la solución óptima para cualquier función cuadrática de estimación de errores [Einicke09] [Carmi+10].

En multitud de aplicaciones, con el fin de controlar un sistema dinámico, se necesita saber qué se está haciendo en primer lugar o que se realiza primero. Para estas aplicaciones, no siempre es posible o deseable medir todas las variables que desea controlar, y el filtro de Kalman proporciona un medio para inferir la información de mediciones indirectas (generalmente ruidosa). Algunas de las cosas increíbles que el filtro de Kalman puede realizar es predecir los posibles estados futuros de los sistemas dinámicos que las personas probablemente no puedan controlar.

A continuación se presentan algunas ventajas del filtro Kalman, comparando con otros filtros famosos como, por ejemplo, el filtro de Wiener [Wiener49] y otros.

- La implementación de este filtro computacionalmente no es una novedad con respecto a otros filtros pero sí es una característica importante que se pueda implementar con una mayor precisión.
- El carácter estacionario del filtro de Kalman no es un inconveniente para problemas deterministas o procesos aleatorios. Existen numerosas aplicaciones de gran relevancia en las que utilizan el filtro para procesos estocásticos no estacionarios.
- El filtro de Kalman es compatible con la formulación de espacio de estados de sistemas dinámicos de control óptimo. Este lo hace muy adecuado para los procesos de estimación y control de dichos sistemas.
- Desde un punto de vista matemático, el filtro de Kalman utiliza métodos de decisión basados en la estadística para la detección de información necesaria y el propio filtro rechaza las medidas anómalas.

2.2.5.1 Proceso de estimación

Antes de realizar la descripción del filtro de Kalman, vamos a comenzar con la descripción matemática de los modelos en el espacio de estados.

Un sistema dinámico puede ser también descrito por un modelo en el espacio de estados. El modelo en el espacio de estados de orden *n* de múltiples entradas y múltiples salidas (MIMO) lineal se describe matemáticamente como sigue [Love+08]:

$$\begin{cases} x_{k+1} = A\,x_k + B\,u_k + \omega_k \\ y_k \;\; = C\,x_k + D\,u_k + v_k \end{cases} \tag{2.22}$$

donde A, B, C y D son matrices $(n \times n)$, $(m \times n)$, $(n \times l)$ y $(m \times l)$ respectivamente (véase Figura 2.8).

En ocasiones, la salida del sistema únicamente depende de los estados del sistema y del ruido, con lo que la ecuación anterior de espacio de estados del sistema se suele expresar con la siguiente notación [Tse+05]:

$$\begin{cases} x_{k+1} = A\,x_k + B\,u_k + \omega_k \\ z_k \;\; = H\,x_k + v_k \end{cases} \tag{2.23}$$

Las variables aleatorias $\omega(k)$ y $v(k)$ representan el *ruido de planta* o *ruido proceso* y *ruido de medida*, respectivamente. El ruido de planta modela el efecto de entradas ruidosas que actúan en los estados mismos, mientras que el ruido de medición modela efectos tales como ruido en los sensores.

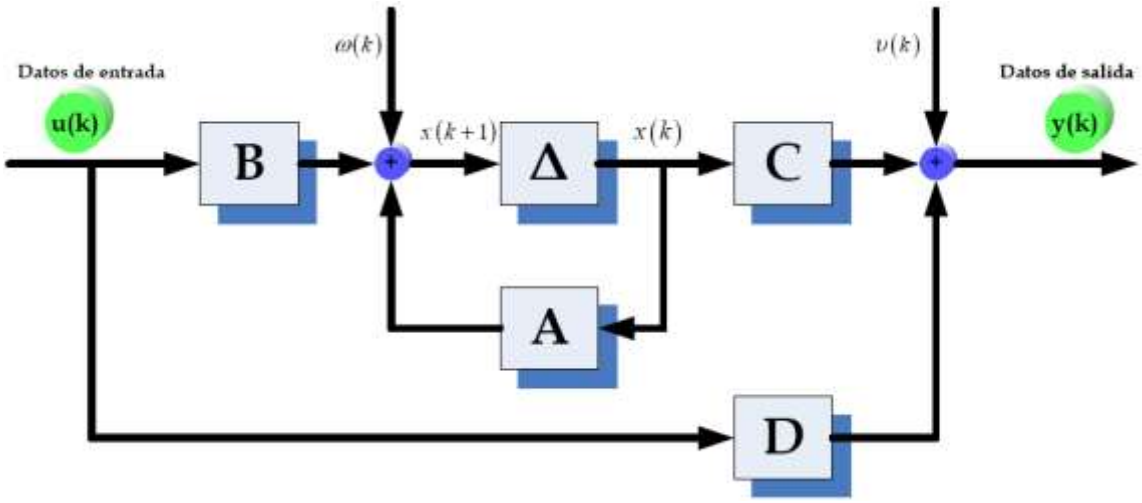

Figura 2.8: Diagrama de estados de un sistema lineal invariante en el tiempo

Fuente: [Ramírez07]

Estos ruidos, son señales de ruido blanco estacionario, de media cero, y no correlacionado entre sí. Esto quiere decir que tienen las siguientes propiedades estadísticas:

$$\begin{cases} E[\omega^T(k)\omega(k)] = Q \\ E[\omega^T(k)\omega(j)] = 0\,,\ k \neq j \\ E[\omega(k)] = 0 \\ E[v^T(k)\omega(j)] = 0\ \ \forall j,k \end{cases} \tag{2.24}$$

$$\begin{cases} E[v^T(k)v(k)] = R \\ E[v^T(k)v(j)] = 0\,,\ k \neq j \\ E[v(k)] = 0 \end{cases} \tag{2.25}$$

donde $E[\cdot]$ denota la media estadística y Q, R son la covarianza de cada ruido.

El estado inicial $x(0)$ es en general desconocido y se modela como una variable aleatoria Gaussiana con media y covarianza conocidas. Las dos secuencias de ruido y el estado inicial se suponen mutuamente independientes [Lim+78].

2.2.5.2 Proceso de corrección

Ahora se definen el parámetro $\hat{x}^-(k)$ *o* $\hat{x}_k^- \in \mathbb{R}^n$ (nótese el superíndice menos de la expresión) como el estado estimado en el paso k a *priori* y $\hat{x}(k)$ *o* $\hat{x}_k \in \mathbb{R}^n$ como

el estado estimado a *posteriori*. Una vez definidos estos dos parámetros estamos en disposición de definir el error estimado a *priori* y a *posteriori*:

$$\begin{cases} e_k^- \equiv x_k - \hat{x}_k^- \\ e_k \equiv x_k - \hat{x}_k \end{cases} \tag{2.26}$$

Con las anteriores definiciones podemos calcular la covarianza del error estimado a *priori*:

$$P_k^- = E[e_k^- e_k^{-T}] \tag{2.27}$$

Y la covarianza del error estimado a *posteriori*:

$$P_k = E[e_k e_k{}^T] \tag{2.28}$$

A la hora de derivar las ecuaciones del filtro Kalman, se comienza por encontrar una ecuación que compute un estado estimado a *posteriori* como una combinación lineal de un estado estimado a *priori* y una ponderación de la diferencia entre la medida actual y una medida de predicción. Esto es:

$$\hat{x}_k = \hat{x}_k^- + K(z_k - H\hat{x}_k^-) \tag{2.29}$$

A la diferencia $(z_k - H\hat{x}_k^-)$ se le llama medida de innovación o residuo [Bay99]. El residuo refleja la diferencia entre la medida predicha y la medida actual. Si el residuo es cero, eso quiere decir que las dos medidas son iguales.

La ganancia K es una matriz de $(n \times m)$ que minimiza el error de covarianza a *posteriori*. Manipulando las ecuaciones (2.28) y (2.29) se puede lograr la expresión de la ganancia K [Gibson+91]:

$$K_k = P_k^- H^T (HP_k^- H^T + R)^{-1} = \frac{P_k^- H^T}{HP_k^- H^T + R} \tag{2.30}$$

Si nos fijamos en la ecuación anterior, podemos observar que si R se aproxima a cero, entonces la ganancia K_k se aproxima a H^{-1}. Por otro lado, si P_k^- tiende a cero, K_k tiende a cero. Todo este resultado se puede interpretar de la siguiente manera: si la covarianza del error de medida R se aproxima a cero, eso quiere decir que podemos confiar más en la medida actual, z_k, y menos en las medidas

predichas, $H\hat{x}_k^-$. Sin embargo, si la covarianza del error estimado a *priori* tiende a cero se puede confiar menos en la medida actual, z_k, y más en las medidas predichas, $H\hat{x}_k^-$ [Gabrea04].

2.2.5.3 Algoritmo de filtro de Kalman

El filtro de Kalman estima un proceso usando una especie de control de la retroalimentación. El filtro estima los estados en un instante de tiempo dado y, posteriormente, obtiene la retroalimentación en forma de medida (ruido). De esta manera, las ecuaciones del filtro de Kalman se pueden separar en dos grupos: ecuaciones de actualización del tiempo (*time update*) y ecuaciones actualización de las medidas (*measurement update*). Las ecuaciones de actualización de tiempo es responsable de proyectar los estados actuales hacia delante, es decir, de *predecir*. El vector de estados actual y la covarianza del error estimado son los que obtiene los estados estimados del próximo paso.

Las ecuaciones de actualización de las medidas son responsable de la retroalimentación del sistema, es decir, de incorporar nuevas medidas al estado a *priori* estimado de cara a obtener un estado estimado a *posteriori*. De esta manera, se puede pensar que las ecuaciones de actualización de las medidas son las ecuaciones de *corrección*.

Hasta este momento, hemos visto algunas ecuaciones del algoritmo del filtro de Kalman pero aún no hemos descrito todas y tampoco las hemos englobado en grupos. Las ecuaciones de predicción del algoritmo de Kalman son:

$$\hat{x}_k^- = A\,\hat{x}_{k-1} + B\,u_k \tag{2.31}$$

$$P_k^- = AP_{k-1}A^T + Q \tag{2.32}$$

Nótese cómo en las ecuaciones anteriores se proyecta el estado del paso $k-1$ al paso k, como se ha mencionado anteriormente. Cabe recordar que A, B son las matrices del sistema y Q es la covarianza del ruido de medida. Recuérdese también que las condiciones iniciales del filtro de Kalman son las mencionadas

anteriormente. Presentemos ahora el bloque de ecuaciones de corrección, que son las ecuaciones (2.29) y (2.30) junto con [Brown+92] [Jacobsen+03]:

$$P_k = (I - K_k H) P_k^- \tag{2.33}$$

La primera tarea durante el proceso de corrección es calcular la ganancia del filtro de Kalman, ecuación (2.30) El siguiente paso es actualizar la medida del sistema, ecuación (2.29) y con ello obtener el estado estimado a posteriori incorporando dicha medida. Finalmente, el último paso es hallar la covarianza del error estimado a posteriori.

Para cada par de pasos de *predicción-corrección*, el proceso se repite con las variables previas estimadas a *posteriori* y con cuyas variables se utilizarán para obtener los nuevos estados estimados a *priori*. Esta medida de recursividad es muy importante de cara a implementar el algoritmo del filtro de Kalman.

El mayor problema del filtro de Kalman es la obtención de las covarianzas del ruido. En la mayoría de las aplicaciones la covarianza del ruido de medida, R, se mide antes de aplicar el filtro. Debido a que inevitablemente tenemos que realizar mediciones de las variables que queremos medir, es posible tener una estimación del error que se obtiene realizando este proceso. Por lo tanto, es relativamente sencillo obtener la covarianza del error de medida (veremos que en el caso de procesado de señal de la voz no se realiza así exactamente).

La determinación de la covarianza del ruido del sistema, Q, por lo general suele ser mucho más complicada ya que no tenemos la posibilidad de medir los procesos que estamos estimando. A veces, se utiliza un modelo del sistema que puede tener resultados aceptables introduciendo incertidumbre en el sistema seleccionando una matriz Q determinada.

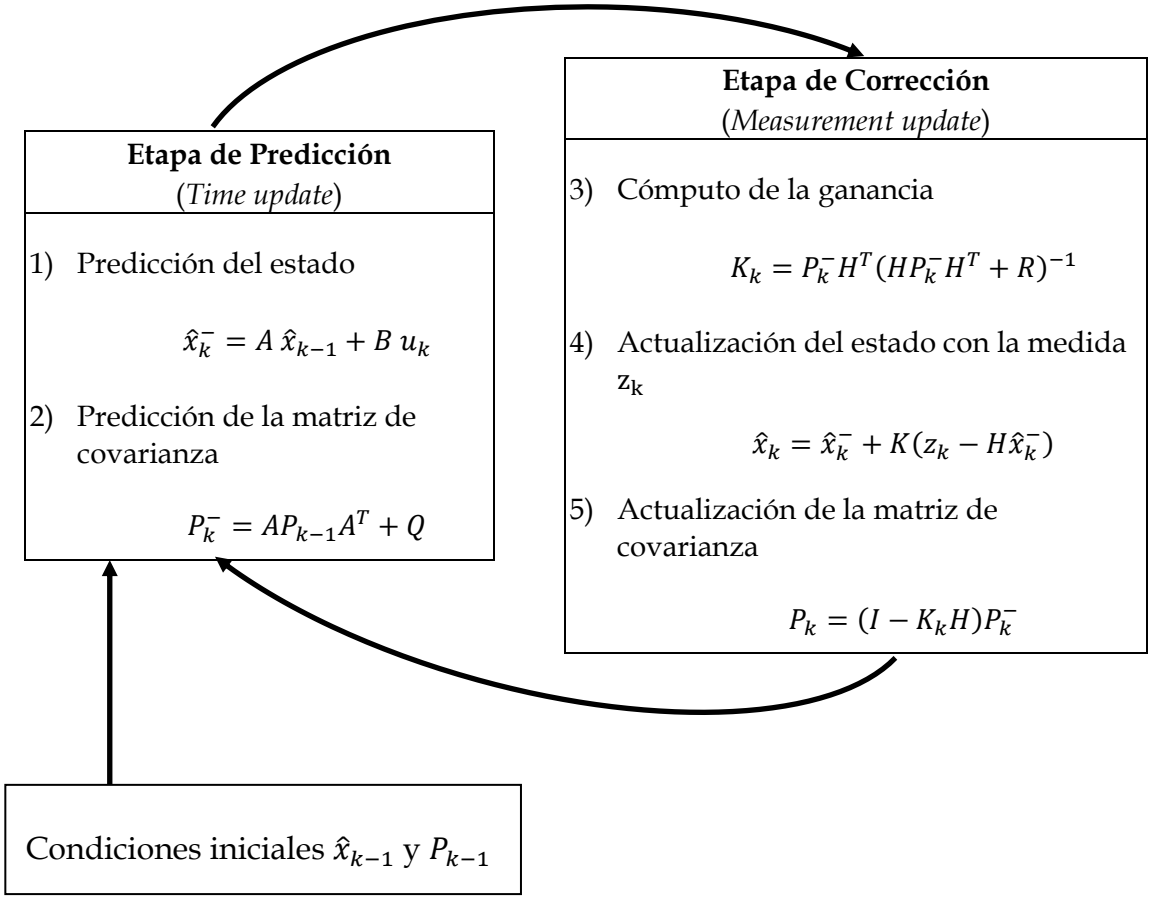

Figura 2.9: Diagrama del algoritmo de Kalman

Existen otras variantes en las que tanto si se poseen bases sólidas para obtener las matrices de covarianza como si no, se pueden estimar *Q, R*. Para ello, se suelen utilizar filtros externos (a veces se utiliza el propio filtro de Kalman) para obtener dichas matrices.

2.2.6 Estabilización de polos

Este algoritmo resume el trabajo realizado en su tesis doctoral la investigadora Begoña García Zapirain [García03]. No obstante, se ve conveniente realizar una pequeña descripción de dicho trabajo. Con lo que se emplaza al lector a consultar dicha tesis para un mayor detalle de este algoritmo de regeneración de la voz esofágica.

Al igual que como se realiza en el desarrollo de esta tesis, el algoritmo de "Estabilización de polos" trabaja también con las voces esofágicas y, concretamente, con la vocal "a".

En líneas generales, este algoritmo calcula la evolución del módulo y fase de cada uno de los formantes de la vocal "a" modificando, si fuera necesario, los polos del tracto vocal. En concreto, este trabajo de investigación se centra en los tres primeros polos del tracto vocal. Estos polos son obtenidos mediante los Coeficientes de Predicción Lineal (LPC) descritos en esta tesis en el apartado 2.2.1. Estabilizando los tres primeros polos del tracto vocal se estabilizan los 3 primeros formantes de la vocal "a".

Este algoritmo, de procesado de señal de la voz esofágica, está dirigido a mejorar la Relación Armónico Ruido (HNR). Al reajustar la posición de los polos dentro del círculo del radio unidad de la transformada Z [Proakis+07], tanto en el módulo como en la fase, se realza los formantes de la vocal "a". Es decir, se eleva el valor del módulo en decibelios de los formantes y esto hace que mejore el parámetro HNR y, por lo tanto, mejore la inteligibilidad de la voz esofágica.

En la Figura 2.10 podemos observar el diagrama detallado del algoritmo de "Estabilización de polos".

El primer bloque de la Figura 2.10 toma como entrada la señal de voz y la codifica según el modelo de coeficientes de predicción lineal (LPC) para calcular los polos de la señal que posteriormente serán analizados y transformados [García+08b].

En el siguiente bloque del diagrama se calcula el módulo y la fase de los tres primeros polos. Si se observa la evolución de dichos polos se aprecia claramente la desproporcionada variación de los mismos. Ésta es debida a que tanto a que el valor del módulo como la fase oscila mucho y experimenta cambios excesivos en su valor.

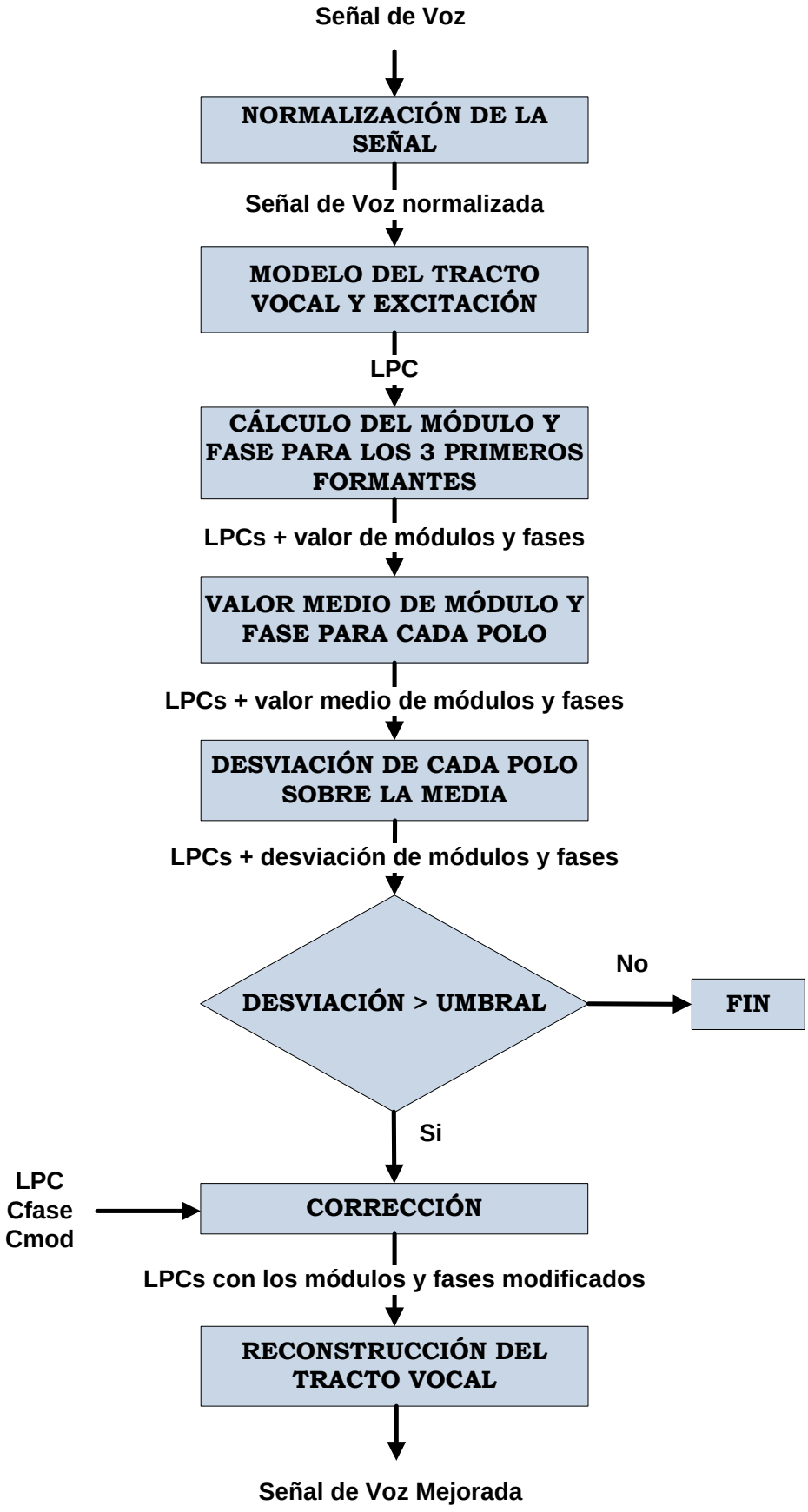

Figura 2.10: Diagrama de bloques de la estabilización de polos

Una vez obtenidos estos datos se podrá trabajar en el procesado de la señal ya que ésta será la base del algoritmo de regeneración de la voz esofágica. Para ello, será necesario reajustar los polos que es la aportación principal de este algoritmo ya que implementará la modificación que permitirá la mejora de la Relación armónicos ruido (HNR).

Después de calcular el módulo y la fase de los tres primeros polos se calcula el valor medio del módulo y la fase, para cada uno de los polos. Esto nos sirve para realizar el siguiente bloque del diagrama de "Estabilización de los polos" que es el de calcular la "desviación de los polos sobre la media".

Posteriormente, se intenta limitar la variación de dichos formantes. Para ello, se mira a ver si la desviación de los polos sobre la media supera un umbral establecido empíricamente. Si este fuera el caso, se limita la variación de los formantes para que no sea excesiva y, como consecuencia de ello, se modifican los tres primeros formantes de la señal vocal, que se identifican con las tres primeras frecuencias de los polos.

Para realizar este proceso, para cada una de las tramas de las que se han calculado los coeficientes, se obtendrán las raíces de los coeficientes de predicción lineal y, de cada raíz, se calculará la frecuencia de cada uno de los polos del sistema como ya se ha comentado.

Al igual que se realiza en el algoritmo del filtrado de Kalman, el orden de los coeficientes de predicción lineal que se emplea es de 14. Por lo tanto, habrá 14 polos, los cuales son complejos conjugados por lo que tendremos 7 polos para las frecuencias positivas y 7 para las frecuencias negativas. Para el proceso de este algoritmo sólo se escogerán los polos de las frecuencias positivas. Para cada trama se utilizarán los 7 polos de frecuencia positiva: de estos siete polos se calculará qué polos corresponden con los 3 de menor frecuencia. Las frecuencias de estos 3 polos corresponden con las frecuencias de los tres formantes principales, que son los formantes que queremos modificar.

A la función de corrección se le pasa como parámetros de entrada la matriz de coeficientes de predicción lineal así como los factores de corrección de frecuencia y de módulo que se quieren aplicar a los formantes. Como resultado se obtienen una matriz con los coeficientes de predicción lineal modificados.

Esta función detecta, para cada instante de tiempo, cuáles son los tres principales formantes del fonema y procura que la evolución de cada uno de dichos formantes, tanto en módulo como en fase, esté comprendida dentro de unos márgenes apropiados. Este proceso de corrección de los formantes se realiza sobre los coeficientes de predicción lineal calculados en un bloque anterior. Debido a esto, a la función que implementa este bloque se le debe introducir los coeficientes de predicción lineal y, además, también se le tendrá que especificar cuáles son los factores de corrección que deberá emplear tanto para corregir la fase como para corregir el módulo. Entre todas las pruebas realizadas en la tesis de Begoña García Zapirain, se ha escogido los valores de las constantes que más mejoran el HNR. En concreto, se ha escogido como constante de fase, Cfase = 0'4, y como constante de módulo, Cmod = 0'3.

Una vez conocida la desviación máxima que sufre cada uno de los formantes tanto en la frecuencia como en el módulo, se realiza un proceso para corregir esta desviación que sufren respecto al valor medio. La corrección de la fase se lleva a cabo mediante la siguiente ecuación:

$$\text{AnguloModif} = \text{Angulo} - (\text{Cfase} * (\text{Angulo} - \text{MediaAngulo})) \quad (2.34)$$

Cada valor del ángulo "AnguloModif" se corrige aplicando la fórmula anterior y, por lo tanto, varía la frecuencia original a la que se encuentra el formante. Como se puede comprobar, el efecto de aplicar dicha fórmula es reducir la diferencia del valor del formante respecto a la media del mismo.

La corrección del módulo se realizará mediante la siguiente fórmula ecuación:

$$\text{ModuloModif} = \text{Modulo} + ((1 - \text{Modulo}) * \text{Cmod}) \quad (2.35)$$

Cada valor de módulo de cada uno de los formantes se modifica haciendo que la variación del valor del módulo sea menor por lo que el formante tendrá un valor de módulo más constante a lo largo del tiempo. El valor "ModuloModif" representa el valor del módulo modificado de cada uno de los polos.

Cuando se realiza un cambio en uno de los polos también se le aplica el mismo cambio al complejo conjugado de forma que el valor del polo modificado tenga su correspondiente polo complejo conjugado.

Con los 14 nuevos polos lo que se realiza es el cálculo del polinomio que describe la función de transferencia del tracto vocal, y este polinomio es el que se utiliza como coeficientes de predicción lineal modificados. Estos son los coeficientes que devuelve el bloque de corrección.

Finalmente, será necesaria la reconstrucción de la señal de voz a partir de la nueva función de transferencia después de la modificación realizada sobre los polos. De esta manera, la salida de este bloque es la señal de voz mejorada.

2.3 PARÁMETROS ACÚSTICOS DE LA VOZ

Uno de los objeto de estudio importante de esta tesis son los parámetros acústicos de la voz. La posibilidad de medir lo que se dice y además ver la voz en una gráfica, resulta muy novedoso como método para la implantación y rehabilitación de la voz y el habla. Poder medir, los perfiles acústicos de los principales parámetros de la señal de voz y asociarlos con imágenes que representan lo dicho, ha resultado una alternativa adicional muy estimulante en el campo de la foniatría, la logopedia y la otorrinolaringología.

En este sentido, también en el mundo informático se estudian los parámetros de la voz, encuadrada en el área de tratamiento digital de la voz. La mayoría de los trabajos se dedican a dos ámbitos bien diferenciados. Por un lado, a la codificación [Novorita99] [Paliwal+93], síntesis [Kim+10] [Wei+09] y reconocimiento del habla y de locutor [Baker+09] [Bourlard+98] [Flanagan65].

Los ámbitos de actuación a los que corresponden esas investigaciones preferentemente corresponden a estudios de hombre-máquina. A su vez, se cuentan con algunas principales técnicas de análisis sobre la señal de voz. Los desarrollos en los ámbitos desarrollados anteriormente centran sus estudios en: predicción lineal, cuantificación escalar-vectorial, Modelos ocultos de Markov, redes neuronales, técnicas de modificación de prosodia de la señal de voz etc.

Por otro lado, otros de los ámbitos de análisis de la voz hacen referencia a: el conocimiento relativo de características elementales de la señal de voz, técnicas de análisis básicos como medidas localizadas de la voz: detección de ciclos de la voz [Chen+01] [Hagmüller+06], cruces por cero [Kedem86], autocorrelación [Cheveigné+02], análisis en frecuencia [Dorken+94], estimación de la frecuencia fundamental [Piszczalski+79], predicción lineal [Grant+00] etc.

Sin embargo, en la mayoría de estas investigaciones, las voces con las que se trabaja son laringadas, sanas o con alguna patología [DeBonis+08] [Gail11] quedando las voces esofágicas fuera del objeto de estudio de los principales equipos de investigación que trabajan sobre estos temas.

Los objetivos de esta tesis son caracterizar y mejorar la voz esofágica. Para ello, es necesario medir los parámetros acústicos de la voz. Los parámetros típicos que se miden en las señales acústicas son la frecuencia fundamental o *pitch* (F_0), perturbación de la frecuencia de ciclo a ciclo o *jitter*, perturbación de la amplitud de la voz ciclo a ciclo o *shimmer* y medidas del ruido, como por ejemplo, *Harmonic to Noise Ratio* (HNR) y/o *Signal to Noise Ratio* (SNR) [Moran+06]. Todos estos parámetros caracterizan la voz de forma objetiva. Hay otros parámetros que caracterizan la voz de forma subjetiva [Flipsen06]. En esta tesis se han utilizado los parámetros objetivos para caracterizar la voz.

Por lo general, los estudios de análisis acústico de la voz se suelen realizar sobre vocales sostenidas y, entre ellas, la más común es la vocal /a/ [Gonzalez+02].

2.3.1 Frecuencia Fundamental o *Pitch*

El pitch, o frecuencia fundamental de la voz, es una de las propiedades del sonido o del tono musical que representa la frecuencia percibida [Gibiat88] [Doval+91]. Debido a esta pseudo-periodicidad natural de la voz sonora, existen pequeñas variaciones en los picos de la voz que hacen que cambie su frecuencia, F_i. Es decir, el pitch se calcula obteniendo los segmentos de la voz donde se repite la onda.

Las marcas donde se repite la onda se llaman "marcas de pitch", "ciclos de pitch" o "épocas", [Baken+00] [Doval+93] [Feijoo+90] [Gonzalez+02] [Moran+06]. Estas marcas de pitch están relacionadas con el ciclo glotal. En la Tabla 2.1 se pueden ver distintas medidas del pitch.

Tabla 2.1: Medidas relacionadas con el pitch

Descripción	Método de cálculo
Media del pitch (Mean F_0, F_0_av)	$\bar{F} = \frac{\sum_{i=1}^{N} F_i}{N}$
Máximo valor detectado del pitch (Maximum F_0, F_0_hi)	$F_0\_hi = max(F_i)$
Mínimo valor detectado del pitch (Minimum F_0, F_0_lo)	$F_0\_lo = min(F_i)$
Desviación estándar de F_0 (F_{sd})	$F_{sd} = \frac{\sum_{i=1}^{N}(F_i - \bar{F})^2}{N-1}$
Rango de frecuencia fonatoria (Phonatoy Frequency Range, PFR)	$PFR = \frac{\log\left(\frac{F_0\_hi}{F_0\_lo}\right)}{\log(2)} \times 12$

Fuente: [Moran+06]

Además de la manera de medir la señal acústica, es fundamental señalar los rangos de normalidad de la voz sana. Los rangos que se presentan a continuación son para personas de ambos géneros y para personas de distintas edades [Gonzalez+02]. Estos son los rangos de normalidad de la voz sana o laringada por género:

Tabla 2.2: Rangos de normalidad del pitch

Descripción	Rango	Media	Rango	Media
	Varones		Mujeres	
Media del pitch (Mean F_0, F_0_av) (Hz)	83 - 153	120	158 - 274	200
Máximo valor detectado del pitch (F_0_hi) (Hz)	85 - 167	125	166 - 303	213
Mínimo valor detectado del pitch (F_0_lo) (Hz)	81 - 148	115	70 - 263	186
Desviación estándar de F_0 (F_{sd}) (Hz)	0,73 - 4,46	1,35	1,16 - 68,21	3,34
Rango de frecuencia fonatoria (Phonatoy Frequency Range, PFR) (Semitonos)	1 – 7	2,48	2 - 21	2,57

Fuente: (Gónzalez, Cervera, & Miralles, 2002)

Otros trabajos de investigación presentan unos niveles de pitch medios inferiores mayores a los mostrados en esta tabla. Concretamente, en el trabajo [Šiupšinskienė03] se presentan unos rangos de normalidad medios de mínimo de pitch para los hombres de 111,7 Hz y de 212,4 Hz para las mujeres. Otros autores [Baken+00] diferencian la frecuencia fundamental no sólo en hombres y mujeres sino que dependiendo de la edad. En caso de los hombres, cuyo pitch es menor que el de las mujeres, el mínimo pitch está en los 102 Hz, algo mayor que el mostrado en la Tabla 2.1.

Tabla 2.3: Medidas de perturbación del pitch, o jitter

Descripción	Método de cálculo
Media del Jitter. Mean Absolute Jitter (jitta o MAJ) (Hz)	$Jitta = \frac{\sum_{i=1}^{N-1}\lvert F_{i+1} - F_i\rvert}{N-1}$
Jitter (%) (Jitt)	$Jitt = \frac{\frac{\sum_{i=1}^{N-1}\lvert F_{i+1} - F_i\rvert}{N-1}}{\frac{\sum_{i=1}^{N} F_i}{N}} = \frac{Jitta}{\bar{F}}$
Perturbación media relativa (Relative Average Perturbation; Smoothed over 3 pitch periods, RAP)	$RAP = \frac{\frac{1}{N-2}\sum_{i=2}^{N-1}\left\lvert \frac{F_{i-1}+F_i+F_{i-1}}{3} - F_i\right\rvert}{\bar{F}} \times 100$
Coeficiente de perturbación del pitch (Pitch Perturbation Quotient; Smoothed over 5 pitch periods, PPQ_5)	$PPQ_5 = \frac{\frac{1}{N-4}\sum_{i=3}^{N-2}\left\lvert \frac{\sum_{k=i-2}^{k=i+2} F_k}{5} - F_i\right\rvert}{\bar{F}} \times 100$
Coeficiente de perturbación del pitch (Pitch Perturbation Quotient; Smoothed over 55 pitch periods, PPQ_55)	$PPQ_{55} = \frac{\frac{1}{N-54}\sum_{i=28}^{N-27}\left\lvert \frac{\sum_{k=i-27}^{k=i+27} F_k}{55} - F_i\right\rvert}{\bar{F}} \times 100$
Factor de perturbación del pitch (Pitch Perturbation Factor, PPF)	$* PPF = \frac{N_{P \geq threshold}}{N_{Voice}} \times 100$
Factor de perturbación direccional (Directional Perturbation Factor, DPF)	$** DPF = \frac{N_{\Delta\pm}}{N_{Voice}} \times 100$

* Donde N_P son periodos de pitch a lo largo del tiempo que tenga más de 0.5ms de magnitud

** Donde $N_{\Delta\pm}$ son periodos de pitch a lo largo del tiempo donde existe un cambio de signo.

Fuente: [Moran+06]

2.3.2 Perturbación de la Frecuencia Fundamental o *Jitter*

El Jitter es un parámetro que representa la variación de la frecuencia fundamental o pitch [Gerratt05]. De hecho, los parámetros de perturbación de la frecuencia fundamental están representados por el jitter. Sin embargo, existe una gran cantidad de parámetros asociados que representan la misma perturbación a lo largo de un tiempo definido de la voz. En la Tabla 2.3 podemos apreciar estos parámetros.

Tabla 2.4: Rangos de normalidad de la perturbación del pitch, o jitter

Descripción	Rango	Media	Rango	Media
	Varones		Mujeres	
Media del Jitter. Mean Absolute Jitter (jitta o MAJ) (Hz)	20,57 - 167,39	57,53	5,84 – 192,23	32,92
Jitter (%) (Jitt)	0,25 - 2,14	0,68	0,15 – 3,83	0,94
Perturbación media relativa (RAP)	0,14 - 1,31	0,39	0,07 – 2,28	0,57
Coeficiente de perturbación del pitch (PPQ_5)	0,15 – 1,21	0,40	0,08 - 2,26	0,55
Coeficiente de perturbación del pitch (Smoothed over 55 pitch periods, PPQ_55)	0,38 – 1,52	0,68	0,30 – 3,61	0,75

Fuente: (Gónzalez, Cervera, & Miralles, 2002)

Al igual que con el pitch se analizan los rangos de normalidad de la voz sana o laringada tanto de varones como de mujeres. En esta ocasión se presentan los siguientes parámetros de normalidad: jitta, jitt, RAP, PPQ_5 y PPQ_55 (Tabla 2.4).

2.3.3 Perturbación de la Amplitud en los Periodos de Pitch o *Shimmer*

El shimmer es un parámetro que representa la perturbación de la amplitud de la señal de voz en los instantes de pitch. La voz que se produce en las cuerdas vocales es prácticamente constante en su amplitud, por lo tanto, un aumento del valor de shimmer puede implicar un síntoma de un trastorno de la voz. En la Tabla 2.5 podemos ver los parámetros relacionados con la perturbación de la amplitud.

Tabla 2.5: Medidas de perturbación de la amplitud, o shimmer

Descripción	Método de cálculo
Media de la amplitud (Mean Amplitud A_0, A_0_av)	$\bar{A} = \frac{\sum_{i=1}^{N} A_i}{N}$
Máximo valor detectado de amplitud (Maximum A_0, A_0_hi)	$A_0\_hi = max(A_i)$
Mínimo valor detectado de amplitud (Minimum A_0, A_0_lo)	$A_0\_lo = min(A_i)$
Desviación estándar de A_0 (A_{sd})	$A_{sd} = \frac{\sum_{i=1}^{N}(A_i - \bar{A})^2}{N-1}$
Media del Shimmer. Mean Absolute Shimmer (MAS)	$MAS = \frac{\sum_{i=1}^{N-1}\lvert A_{i+1} - A_i\rvert}{N-1}$
Shimmer (%) (Shim)	$Shim\ (\%) = \frac{\frac{\sum_{i=1}^{N-1}\lvert A_{i+1} - A_i\rvert}{N-1}}{\frac{\sum_{i=1}^{N} A_i}{N}} = \frac{MAS}{\bar{A}}$
Shimmer en decibelios (ShdB)	$Shim\ (dB) = \frac{\sum_{i=1}^{N-1} 20 \times \log\left(\frac{A_{i+1}}{A_i}\right)}{N-1}$
Perturbación media relativa de la Amplitud (Amplitud Relative Average Perturbation; Smoothed over 3 pitch periods, ARP)	$ARP = \frac{\frac{1}{N-2}\sum_{i=2}^{N-1}\left\lvert\frac{A_{i-1} + A_i + A_{i-1}}{3} - A_i\right\rvert}{\bar{A}} \times 100$
Coeficiente de perturbación de la Amplitud (Amplitud Perturbation Quotient; Smoothed over 5 pitch periods, APQ_5)	$APQ_5 = \frac{\frac{1}{N-4}\sum_{i=3}^{N-2}\left\lvert\frac{\sum_{k=i-2}^{k=i+2} A_k}{5} - A_i\right\rvert}{\bar{A}} \times 100$
Coeficiente de perturbación de la Amplitud (Amplitud Perturbation Quotient; Smoothed over 55 pitch periods, APQ_55)	$APQ_{55} = \frac{\frac{1}{N-54}\sum_{i=28}^{N-27}\left\lvert\frac{\sum_{k=i-27}^{k=i+27} A_k}{55} - A_i\right\rvert}{\bar{A}} \times 100$
Factor de perturbación de la Amplitud (Amplitud Perturbation Factor, APF)	$* APF = \frac{N_{P \geq threshold}}{N_{Voice}} \times 100$
Factor de perturbación direccional de la Amplitud (Amplitud Perturbation Factor, ADPF)	$** ADPF = \frac{N_{\Delta\pm}}{N_{Voice}} \times 100$

* Donde N_P son periodos de pitch a lo largo del tiempo que tenga más de 0.5ms de magnitud

** Donde $N_{\Delta\pm}$ son periodos de pitch a lo largo del tiempo donde existe un cambio de signo.

Fuente: [Moran+06]

En la Tabla 2.6 se presentan los rangos de normalidad para el shimmer. Entre los parámetros que se presentan son: el shimmer en decibelios (ShdB), Shim, el APQ_5, APQ_55.

Tabla 2.6: Rangos de normalidad de perturbación de la amplitud, o shimmer

Descripción	Rango	Media	Rango	Media
	Varones		Mujeres	
Shimmer en decibelios (ShdB)	0,11 - 0,74	0,33	0,11 -0,91	0,34
Shimmer (%) (Shim)	1,33 – 8,33	3,82	1,33 – 9,58	3,89
Coeficiente de perturbación de la Amplitud (APQ_5)	1,34 – 7,06	3,06	1,09 – 5,97	2,87
Coeficiente de perturbación de la Amplitud (APQ_55)	2,69 – 9,31	4,87	1,87 – 11,06	5,13

Fuente: (Gónzalez, Cervera, & Miralles, 2002)

2.3.4 Ruido de la Señal de Voz

Existen varios parámetros para medir el ruido de una señal de voz. Entre ellos, podemos destacar el "*Harmonic to Noise Ratio*" (*HNR*), "*Signal to Noise Ratio*" (*SNR*), "*Normalized Noise Energy" (NNE), "Glottal-to-Noise Excitation Ratio" (GNE)* etc.

2.3.4.1 Harmonic to Noise Ratio (HNR)

El parámetro "*Harmonic to Noise Ratio*" (*HNR*) es un indicador comúnmente utilizado en el análisis automático de la señal con el fin de obtener el nivel de ruido de la señal de voz. Ésta es también una forma de distinguir las señales de

voz sanas de las que no lo son [Kumara07]. Las señales con mucho ruido son indicadores de inestabilidad o irregularidad de la voz [Girin06] [Murphy+08].

El *Harmonics-to-noise ratio* (*HNR*) se calcula como la media de la amplitud de los componentes puramente periódicos de la señal divido por la media la amplitud de los componentes de ruido. El HNR es expresado generalmente en escala logarítmica, es decir, en decibelios.

Existen varios métodos para calcular el HNR de la voz. Entre los más conocidos y utilizados son los desarrollados por Yumoto [Yumoto+82] y Boersma [Boersma93]:

- Algoritmo de Yumoto para el HNR [Yumoto+82]:

Este algoritmo está realizado en el dominio temporal el cual trabajo con una media del patrón de pitch. El método asume que una onda acústica de una vocal sostenida tiene dos componentes: una periódica que es similar de ciclo a ciclo y un componente ruido cuya distribución de amplitud tiene como media cero. Cada pulso de pitch se representa por medio una señal, $s(t)$, que se mantiene constante durante la señal para todos los N ciclos. Por otro lado, tenemos los términos de ruido de la señal: $e_i(t)$ donde i es el índice de 1 a N de los ciclos de pitch de la señal. De esta manera, la señal de voz se puede representar [Qi+97]:

$$x_i(t) = s(t) + e_i(t) \tag{2.36}$$

La cuestión ahora es construir un patrón de pulso de esta señal. Esto se realiza haciendo la media de todos los pulsos:

$$\bar{x}(t) = \frac{1}{N}\sum_{i=1}^{N} x_i(t) = s(t) + \frac{1}{N}\sum_{i=1}^{N} e_i(t) \tag{2.37}$$

Una vez descritas las señales el HNR se define de la siguiente manera:

$$HNR_{Yumoto} = \frac{N \cdot E\left[\left(\bar{x}(t)\right)^2\right]}{\sum_{i=0}^{N}\left[\left(x_i(t) - \bar{x}(t)\right)^2\right]} \tag{2.38}$$

La limitación del algoritmo es que necesita un número de ciclos suficientemente elevado, entre 30 y 50 ciclos. Por otro lado, el jitter y del shimmer tienen influencia sobre el valor del HNR. Para solucionar estas limitaciones, en 2005 se presenta un trabajo de investigación por Carlos A. Ferrer [Ferrer+05]. La solución propuesta en este trabajo para la medida del HNR es la siguiente [Ferrer+06]:

$$HNR = \frac{N-1}{N} HNR_{Yumoto} - \frac{1}{N} \tag{2.39}$$

- Algoritmo de Boersma para el HNR [Boersma93]

El algoritmo de Boersma, al igual que el de Yumoto, es un método que se circunscribe al dominio temporal. En él se utiliza la función de autocorrelación enventanada de la señal de voz para determinar el pitch y, posteriormente, el HNR. La función de autocorrelación señal se define:

$$r_x(\tau) \triangleq \int x(t)x(t+\tau)dt \tag{2.40}$$

El periodo fundamental T_0 está definido como el valor de τ correspondiente al valor máximo (el índice cero está excluido) de la función de autocorrelación. La energía de la señal de voz enventanada es valor de la función de autocorrelación con el índice cero.

$$r_x(0) = r_p(0) + r_{ap}(0) \tag{2.41}$$

donde $r_p(0)$ y $r_{ap}(0)$ son las energías de la componente periódica y aperiódica de la señal, respectivamente. La función normalizada de la función de autocorrelación se define como:

$$r_x'(\tau) = \frac{r_x(\tau)}{r_x(0)} \tag{2.42}$$

Dada la periodicidad de la componente periódica de la función de autocorrelación y, asumiendo que el ruido no está correlacionado, es decir, que es como si fuera un ruido blanco, la energía de la componente periódica es:

$$r_p'(0) = r_p'(T_0) = r_x'(T_0) \tag{2.43}$$

entonces la energía de la componente no periódica es:

$$r_{ap}'(0) = 1 - r_p'(0) = 1 - r_x'(T_0) \quad (2.44)$$

Por lo tanto, una vez definidas las señales, el HNR se define como [Ferrand02]:

$$HNR \triangleq \frac{r_p'(0)}{r_{ap}'(0)} \quad (2.45)$$

Aunque este algoritmo esté basado en que haya una distribución de ruido no correlacionada en la componente aperiódica, se puede demostrar que funciona relativamente bien para variaciones del jitter.

El rango de normalidad para el HNR en voces sanas se presenta en la Tabla 2.7:

Tabla 2.7: Rangos de normalidad para el HNR

Descripción	Rango	Media	Rango	Media
	Varones		Mujeres	
HNR	4,34 – 10	7,14	4,16 -16,67	7,69

Fuente: (Gónzalez, Cervera, & Miralles, 2002)

Estos algoritmos trabajan en el dominio temporal. Existen otros algoritmos que trabajan en el dominio de frecuencia. A continuación se exponen algunos de ellos:

- Algoritmo de Kojima para el HNR

Este algoritmo [Kojima+80] utiliza la Transformada Discreta de Fourier. Tiene la limitación que es necesario saber el número exacto de pulsos de pitch de la señal. En trabajos posteriores se han realizado correcciones de este algoritmo pero no ha tenido la plena aceptación de la comunidad internacional.

- Estimación del HNR basado en el Cepstrum (Rahmonic)

Estas investigaciones están basadas en el Cepstrum para estimar el HNR. Este tema está publicado por Krom en 1994 [Krom94] y posteriormente fue

modificado por P.J. Murphy en algunos artículos de investigación [Murphy06] [Murphy+07a]. El Cepstrum es la Transformada Inversa de Fourier del espectro de una señal (en nuestro caso la de voz) en escala logarítmica. El Cepstrum contiene información en diferentes bandas de frecuencia. Se expresa en forma de números complejos con lo que tiene parte imaginaria y real. A los picos locales del espectro se les llaman *Rhamonics*. La amplitud del primer Rhamonic, R_1, está relacionado con la calidad de la voz y puede ser utilizado para obtener el HNR de la señal de voz [Murphy+07b].

2.3.4.2 Signal to Noise Ratio

La relación entre la señal y el ruido (Signal to Noise Ratio) es un parámetro más genérico que HNR con respecto al ámbito en el que se utiliza. El SNR se utiliza mayormente para medir todo tipo de señales circuitos eléctricos, aunque también se usa para medir: señales biomédicas en las que se incluyen el ruido de las imágenes, niveles de isotopos etc. Básicamente lo que mide es la calidad de la señal con respecto al ruido de fondo.

La relación entre la señal y el ruido se define así:

$$SNR = \frac{P_{Signal}}{P_{Noise}} \tag{2.46}$$

Donde P es la potencia de media [Qi+99]. Esta medida también se suele expresar en decibelios, con lo que la expresión quedaría de la siguiente forma:

$$SNR_{dB} = 10\, log_{10}\left(\frac{P_{Signal}}{P_{Noise}}\right) = P_{Signal,dB} - P_{Noise,dB} \tag{2.47}$$

Ambas señales deben ser medidas de forma equivalente en la voz y entre los mismos anchos de banda.

2.3.4.3 Normalized Noise Energy (NNE)

La energía del ruido normalizada, Normalized Noise Energy (NNE), es un parámetro que describe el ruido aditivo en una señal (Kasuya, Ogawa, Mazuma,

& Ebihara, 1986). La definición difiere muy poco del Harmonic to Noise Ratio (HNR) dependiendo de los autores. El parámetro Normalized Noise Energy (NNE) es la relación entre la energía de ruido y la energía total de la señal. Existen diferentes implementaciones para el cálculo de este parámetro. Algunos autores, [Michaelis+98],presentan la opción de obtener la energía de ruido directamente entre los armónicos de su espectro, asumiendo dentro de un armónico la energía de ruido como el valor medio de las mínimas adyacentes en el espectro.

2.3.4.4 Glottal-to-Noise Excitation Ratio (GNE)

El ratio entre la excitación glotal y el ruido, Glottal-to-Noise Excitation Ratio (GNE), es un parámetro que describe el ruido aditivo en una señal [Michaelis+97]. El cálculo de este parámetro está basado en la correlación de la envolvente de la transformada de Hilbert entre diferentes canales de frecuencia. Si una señal no tiene ruido las transformadas de Hilbert de las envolventes en diferentes frecuencias son muy similares y, por lo tanto, su correlación es muy elevada. Si por el contrario, alguna de las envolventes tuviera ruido aditivo no correlado, no habría correlación entre las diferentes bandas de frecuencia. Se trata de una buena alternativa al uso de HNR por hacerlo prácticamente independiente del jitter y shimmer.

BASE DE DATOS

3. BASE DE DATOS

"La pérdida del habla es la principal consecuencia de la laringectomía total. Cuando despertamos de la anestesia la primera reacción es intentar hablar. A veces por pura intuición y otras para ver si es cierto que no podemos"

(Asociación Vizcaína de Laringectomizados)

Para recopilar una base de datos de voces esofágicas ha sido necesario contactar con una asociación local de laringectomizados: La Asociación Vizcaína de Laringetomizados y mutilados de la voz (AVL). Esta asociación ha prestado sus voces para la investigación gentilmente, es decir, sin ánimo de lucro.

La Asociación Vizcaína de Laringectomizados y Mutilados de la Voz, lleva desde 1988 trabajando con el objetivo de ayudar a las personas afectadas, es decir, aquellas que a raíz de un cáncer de laringe han sufrido una laringectomía parcial o total. En la Figura 3.1 se presenta el presidente de la asociación. El objetivo principal de la asociación es el de la recuperación del habla de las personas que han sufrido una Laringectomía total, es decir, que nadie se quede sin la oportunidad de recuperar el habla. Para ello, se imparten clases foniátricas diarias (excepto fines de semana). Para estas clases disponen de varios Monitores-Educadores, previamente instruidos a tal efecto en centros donde existen escuelas de monitores, como la Asociación de Laringectomizados de León. Todos los años se envían personas laringectomizadas a estos centros para después poder enseñar en la Asociación Vizcaína de Laringectomizados.

Figura 3.1: Presidente de la Asociación Vizcaína de Laringectomizados

Fuente: http://asbila.jimdo.com/

Otros objetivos de la Asociación Vizcaína de Laringectomizados son:

- Dar apoyo moral tras la intervención quirúrgica, realizada por compañeros que han vivido la misma operación y situación.
- Disponer de la mayor información posible para asesorar de forma, tanto individual como colectiva, sobre aspectos relacionados con la situación sanitaria, familiar, psíquica, etc.
- Dinamizar y reagrupar el colectivo a través de actividades sociales y recreativas.
- Mitigar con una educación especial el trauma que padece el recién operado y asistirlo en todo lo necesario.
- Organizar conferencias y charlas llevadas a cabo por personas especializadas de nuestra Provincia.

Desde esta Asociación se intenta recopilar toda la información existente, relativa a la higiene, cuidados sanitarios, problemas psicológicos, familiares, económicos, etc. Esta información se facilita mediante entrevistas individuales a los asociados, familiares y otros laringectomizados. Con el objeto de aumentar sus conocimientos sobre todo lo relacionado con la laringectomía, asisten a todos los Congresos y reuniones que se celebran en España y en el Extranjero. También se mantienen informados a través de los servicios médicos relacionados con la

laringectomía y similares. De esta manera procuran dar y mantener un servicio de calidad a todos sus asociados.

Como ya se ha comentado en el apartado de Introducción, las personas laringectomizadas deben aprender a hablar de nuevo. En la Asociación Vizcaína de Laringectomizados existen tres niveles de dificultad de aprendizaje del habla. El primer nivel es el nivel más básico de aprendizaje, en el que los afectados aprenden a pronunciar sus primeras palabras. El segundo nivel es el nivel medio y, por último, el tercer nivel es el nivel más avanzado.

Las voces recopiladas para la base de datos de esta investigación son de personas laringectomizadas que tras realizar todo el proceso de aprendizaje del habla, ya han superado el tercer nivel de aprendizaje, es decir, el avanzado. La calidad de la voz esofágica es tan pobre que no es viable trabajar con voces que no estén debidamente entrenadas. Es por ello, sólo se han realizado grabaciones de personas que están en el tercer curso de aprendizaje del habla o aquellas que ya los han realizado.

Los registros de voz que componen la base de datos se han obtenido utilizando un minidisk portable MZR700PC con micrófono incorporado con las siguientes características:

- Grabador de audio con los ajustes correspondientes para los niveles de grabación.
- Regrabable
- Posee auriculares
- Mando a distancia
- Herramienta de repetición del audio
- Sonido pregrabado
- Posibilidad de poner nombre a las pistas de audio
- Dial de selección
- Convertidor de frecuencias de muestreo
- Herramientas de edición

- Buffer para evitar saltos en la reproducción por vibraciones fuertes ("G-Protection By Sony").
- Control de los sonidos graves ("Bass Boost")
- Sincronizado
- Herramientas de edición
- Entrada óptica y de micrófono
- Batería recargable con adaptador AC
- Autonomía de la batería 13 horas
- Accesorios: cable óptico y enlace de cable digital
- Dimensiones: 7,62 cm x 8,10 cm x 2,87 cm
- Peso: 118 gr

El minidisc utiliza un sistema de grabación digital de sonido. La grabación magneto-óptica es un sistema combinado que graba de forma magnética, pero reproduce de forma óptica. Los datos se graban en el disco mediante lo que se conoce como recubrimiento de cambio de fase. Como en un disco compacto, el minidisc almacena la música en pistas.

La frecuencia de muestro recogida en toda la base de datos y la que este grabador minidisk proporciona es la estándar, es decir, de 44,1 kHz.

Las grabaciones se han realizado en el propio local de la Asociación Vizcaína de Laringectomizados, en una sala de reuniones vacía con las mayores condiciones de silencio posibles.

La base de datos está compuesta por más de 300 registros de voces de vocales, palabras y frases completas. Para esta investigación se ha utilizado la vocal sostenida "a". Esta es la vocal que se utiliza por la comunidad científica internacional en numerosos trabajos de investigación [Gonzalez+02] [Kearney04]. De toda la base de datos se han seleccionado 30 voces esofágicas para ser mejoradas con algoritmos de procesado de señal. Otras diez se han escogido para la parametrización de la voz esofágica y han sido comparadas con diez voces

laringadas o "normales". Estas voces normales se han recopilado de voluntarios que han prestado sus voces para esta investigación.

Existe otra base de datos de voces patológicas como la recogida por la compañía Kay Elemetrics: "Disordered Voice Database" [Elemetrics94]. Es una base de datos clínica recopilada por la propia compañía y por los especialistas de patologías del habla en el Massachusetts Eye and Ear Infirmary (MEEI), Boston. La base de datos está provista de muestra de fonemas sostenidos en el tiempo y de voz normalmente hablada. Se grabaron a más de 700 pacientes con diversas patologías de naturaleza orgánica, neurológica, traumática, psicogénica etc. La base de datos comprende voces patológicas de vocales sostenidas como "a", "i" y "u". Solamente del fonema "a" se proveen 650 ficheros de voz. Además, estas vocales se acompañan con muestras de voces sanas de más de 60 personas. Toda la base de datos es recogida con una frecuencia de muestreo de 44,1 kHz con una resolución de 16 bits.

Lamentablemente, no hay registros de voz de habla esofágica. Es por ello, que se ha tenido que desechar esta base de datos y recopilar una propia con ayuda de una asociación de laringectomizados local.

Para completar la base de datos se han realizado grabaciones a 10 personas, con una edad comprendida entre los 60 y 70 años. A cada una de estas personas se les ha realizado 3 grabaciones del fonema "a". Cabe destacar que todas las personas grabadas han sido varones debido a que, como ya se ha comentado durante esta tesis, hay muy pocas mujeres que aprender a hablar voz esofágica por el pudor de tener una voz tan grave.

Esta base de datos ha sido registrada en el "Registro General de la Propiedad Intelectual" en el Departamento de Educación, Política Lingüística y Cultura (Vitoria-Gasteiz, 30/01/2013) con el número de solicitud BI-596-12 y cuya fecha de presentación y efectos es 14/09/2012 [Oleagordia+12].

- **Estructura de la Base de Datos**

La base de datos de voz esofágica para el fonema "a" está estructurada de forma que contiene 30 ficheros en formato de audio con extensión de fichero ".wav". El nombre del fichero viene dado por la siguiente estructura: "a" seguido del número comprendido entre 1 y 30.

Aparte de estas voces, se han utilizado otras 10 voces sanas y 10 esofágicas para la fase de parametrización de la voz del mismo fonema "a".

- **Sistema de acceso a los datos**

Los datos están accesibles en un Compact Disk (CD) que consta de 30 ficheros.

- **Modo de acceso**

El modo de acceso a los datos es directo. Como ya se ha comentado la base de datos consta de 30 ficheros de audio en formato ".wav" a los cuales se accede realizando un doble clic sobre el fichero.

DISEÑO DEL ALGORITMO

4. DISEÑO DEL ALGORITMO

"Un científico debe tomarse la libertad de plantear cualquier cuestión, de dudar de cualquier afirmación, de corregir errores"

Julius Robert Oppenheimer

En este apartado se explicará pormenorizadamente el desarrollo de los algoritmos que validan la hipótesis planteada en esta tesis y cumplen con los objetivos de la misma. El objeto de estos algoritmos es, por una parte, mejorar la calidad de la voz esofágica y, por otra, parametrizar la voz esofágica. El diseño final realizado se ha dividido en dos bloques: uno, **el diseño de alto nivel**, donde se describen los bloques que comprenden el algoritmo a grandes rasgos y, el otro, **el diseño de bajo de nivel**, donde se explican cada uno de los bloques detalladamente. En este último bloque se expondrá específicamente las funcionalidades de cada una bloques expuestos en el diseño de alto nivel.

4.1 DISEÑO DE ALTO NIVEL

El diseño de alto nivel engloba un conjunto de funciones que constituyen la arquitectura del sistema. Como ya se ha comentado anteriormente, en este apartado no se entrará en detalle sobre cada una de las funciones y se reservará esta pormenorización para el diseño de bajo nivel.

En la mayoría de los diseños se realiza una división de las funcionalidades del sistema dando lugar a distintos bloques o etapas.

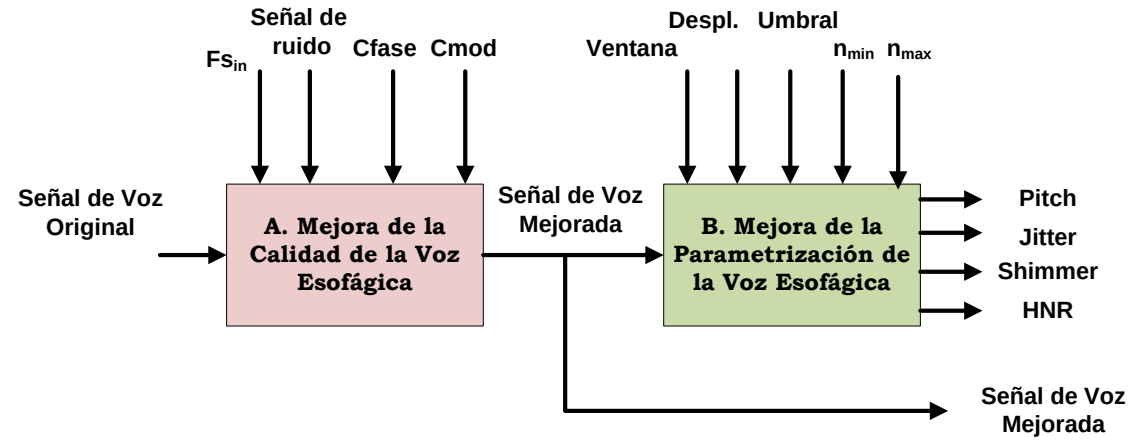

Figura 4.1: Diagrama de bloques del Algoritmo de la Voz Esofágica

Cada bloque o etapa comprende un grupo de funciones del sistema y todos ellos comparten algún elemento común. Cada uno de estos grandes bloques o etapas se puede dividir, a su vez, en más sub-bloques o sub-etapas.

El algoritmo de bloques que se propone para alcanzar los objetivos de esta investigación es el de la Figura 4.1. La entrada del sistema son las señales de voz laringectomizadas descritas en el apartado 3 de este trabajo. El bloque A, "Mejora de la Calidad de la Voz Esofágica", realiza un procesado de la señal mejorando la voz para que se parezca lo máximo posible a una voz laringada. Además de esta señal de voz, este bloque posee tres entradas o "**inputs**" más, las cuales son: la frecuencia de muestreo de la señal de la voz original (Fs_{in}), una señal de ruido de aspiración (ruido en los instantes de silencio de la fonación) y las constantes Cfase y Cmod del algoritmo explicado en el apartado 2.2.4. La señal de ruido mencionada anteriormente es la que se obtiene en los momentos de silencio de la voz esofágica. La salida o "**output**", por tanto, de dicho bloque es una señal de voz mejorada.

El bloque B, "Mejora de la Parametrización de la Voz Esofágica", también tiene como entrada una señal de voz. La entrada de este bloque es la señal de voz mejorada del bloque A. Para la correcta realización del algoritmo es necesario asignar ciertos parámetros dependiendo de la señal de voz de la entrada. Estos parámetros son: ventana, desplazamiento, umbral, n_{min} y n_{max}. La salida de esta

etapa es la medida del pitch, Jitter, Shimmer y HNR. Una vez obtenidas las marcas de la voz mediante el algoritmo propuesto y, de esta manera el pitch (Tabla 2.1), es inmediato obtener el jitter (Tabla 2.3), shimmer (Tabla 2.5) y HNR (apartado 2.3.4).

Este segundo bloque B es necesario para medir de forma automática las mejoras realizadas en el primer bloque A.

A continuación se describen ambos bloques que a su comprenden el diseño de alto nivel de la Figura 4.1 (distinguidos con un código de colores y de letras que se mantendrá a lo largo del capítulo).

4.1.1 Mejora de la Calidad de la Voz Esofágica

El primer bloque es el algoritmo de procesado digital de señal mejora la voz esofágica, "**Mejora de la Calidad de la Voz Esofágica**" (Figura 4.1, Bloque A). En este bloque se describirán una serie de algoritmos que mejoran los parámetros objetivos de la voz esofágica ya descritos en el apartado 2.3 de esta tesis. Este bloque se presenta en la Figura 4.2.

La entrada o "**input**" de esta etapa es una señal de voz esofágica. Estas voces son las obtenidas de la grabación a las propias personas que han sufrido una laringectomía. En concreto, se ha utilizado distintas señales del fonema "a" como entradas de esta etapa. Además, de esta entrada "principal", hay otras tres entradas. La primera, es la frecuencia de muestreo de la señal de entrada, denominada Fs_{in}. La segunda, es una señal de ruido obtenida en los momentos de silencio de la voz. Las otras dos entradas son un par de constantes necesarias para la correcta realización del algoritmo que se detallará en el diseño de bajo nivel.

La salida o "**output**" de este bloque es una señal de voz mejorada. Es decir, una señal de voz cuyas características acústicas se asemejan más que la señal original a las características de una señal de voz laringada. La frecuencia de muestreo de la señal de salida (Fs_{out}) es la misma que la señal de entrada (Fs_{in}).

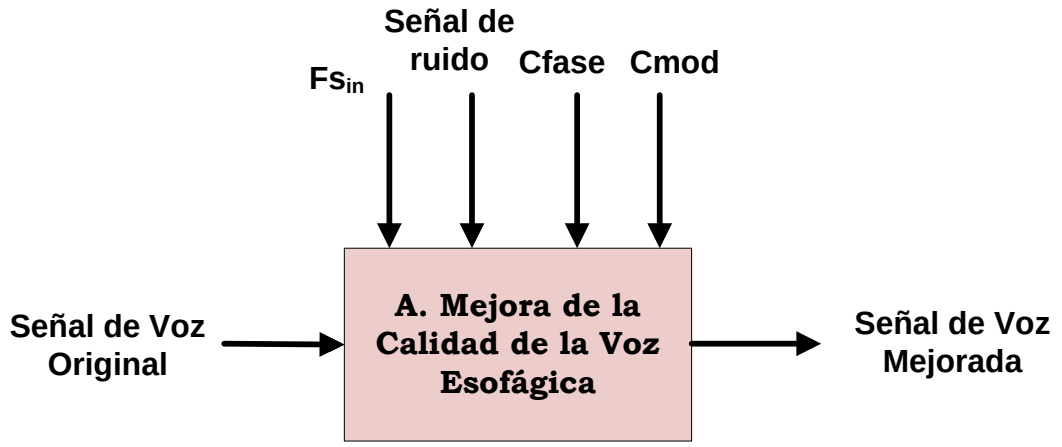

Figura 4.2: Bloque de "Mejora de la Calidad de la Voz Esofágica"

Para la mejora de la calidad de la voz esofágica se han combinado varias técnicas de procesado digital de señal. Se han utilizado tres técnicas de procesado de señal:

- Transformada Wavelet (Discrete Wavelet Transform, DWT)
- Filtro de Kalman
- Estabilización de polos

La primera técnica de procesado está relacionada con la transformada wavelet, la segunda, es la implementación en voz de un filtro de Kalman y, por último, se ha utilizado una estabilización de polos previamente existente para la mejora de las voces esofágicas.

De todos los parámetros descritos en el apartado 2.3, y susceptibles de ser mejorados en esta etapa del algoritmo, se ha hecho especial hincapié en el Harmonic to Noise Ratio (HNR) y en el Shimmer.

4.1.2 Mejora de la Parametrización de la Voz Esofágica

El segundo bloque es el de la "**Mejora de la Parametrización de la Voz Esofágica**" (Figura 4.1, Bloque B). En este bloque consta de un algoritmo que mejora la determinación de los ciclos de la voz que hace posible medir el Pitch de forma adecuada en las voces esofágicas. A partir de aquí, se podrá calcular de

forma efectiva el Jitter, Shimmer y HNR utilizando las ecuaciones descritas en el apartado 2.3. Este bloque se presenta en la Figura 4.3.

La entrada o "**input**" de esta etapa proviene de la señal voz mejorada en el bloque A, descrito anteriormente. Para comprobar el correcto funcionamiento del algoritmo, se ha comparado los resultados con una base de datos de voces sanas o laringadas. Al igual que en el bloque anterior, aquí también se ha tomado como referencia el fonema "a" y se ha caracterizado dicho fonema para voces laringadas y esofágicas, de cara a realizar una comparación de los resultados. Se han medido los parámetros antes mencionados para todo el conjunto de la base de datos. Primeramente, estos parámetros se han medido poniendo las marcas de los ciclos de la voz manualmente con el paquete de software "Multi Dimensional Voice Program", en adelante MDVP, [Deliyeski93] [Nicastri+04] de cara a ser comparados con nuestro algoritmo.

Además de esta entrada, para el correcto funcionamiento del algoritmo es necesario la asignación de ciertos parámetros como son: ventana, desplazamiento, umbral, n_{min} y n_{max}. Estos parámetros se explicarán con detalle en el apartado de "Diseño de bajo nivel".

La salida o "**output**" de este algoritmo son los instantes pitch o marcas de los ciclos de la voz. Como ya se ha comentado anteriormente, con todos y cada uno de los instantes de pitch es inmediato el cálculo de pitch, el cual se puede obtener utilizando las ecuaciones de la tabla 2.1. Por tanto, se compararán los resultados de los parámetros obtenidos para determinar la fiabilidad de la posición de las marcas dadas por el algoritmo. Además, del pitch también es inmediato obtener otros parámetros acústicos como son el jitter, shimmer y el HNR.

Cabe destacar que este algoritmo está diseñado para medir tanto voces sanas como esofágicas.

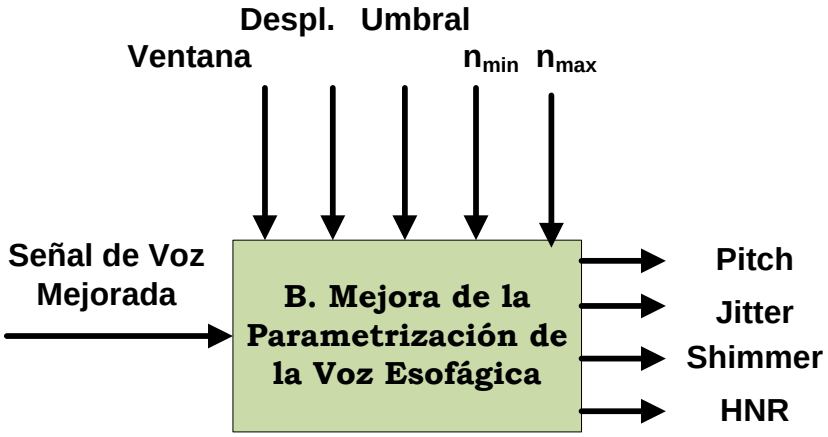

Figura 4.3: Bloque de "Mejora de la Parametrización de la Voz Esofágica"

Este bloque de parametrización alberga algoritmos en el dominio temporal para obtener los instantes de pitch. En algún bloque del mismo se calcula la sonoridad transformando la señal al dominio frecuencial, pero principalmente es un procesado en el tiempo. Se establece una mayor probabilidad de dónde se pueden encontrar los instantes de pitch y se extraen dichas marcas. Después, se discrimina qué tipo de voz es, es decir, o sana o esofágica y finalmente, una vez detectado el tipo de voz, se vuelve a aplicar el algoritmo con unos parámetros establecidos. Para una mayor exactitud del algoritmo se realiza un refinamiento final excluyendo o incluyendo marcas de voz inadvertidas.

Las etapas de este bloque B se pueden enumerar de la siguiente manera:

- Algoritmo Base: algoritmo en el dominio temporal donde se calcula una primera aproximación del pitch.
- Clasificación: se realiza una clasificación de la voz diferenciando las voces sanas de las esofágicas.
- Asignación de parámetros: se asignan una serie de parámetros dependiendo de la voz, sana o esofágica.
- Algoritmo Base: se realiza una segunda pasada del algoritmo con los parámetros asignados del apartado anterior.
- Corrección: se realiza una comprobación de las marcas de la señal obtenidas para detectar posibles errores.

- Cálculo de parámetros: con el vector base de las marcas se obtienen los parámentros pitch, jitter, shimmer y HNR.

Este esquema se puede apreciar en la Figura 4.4:

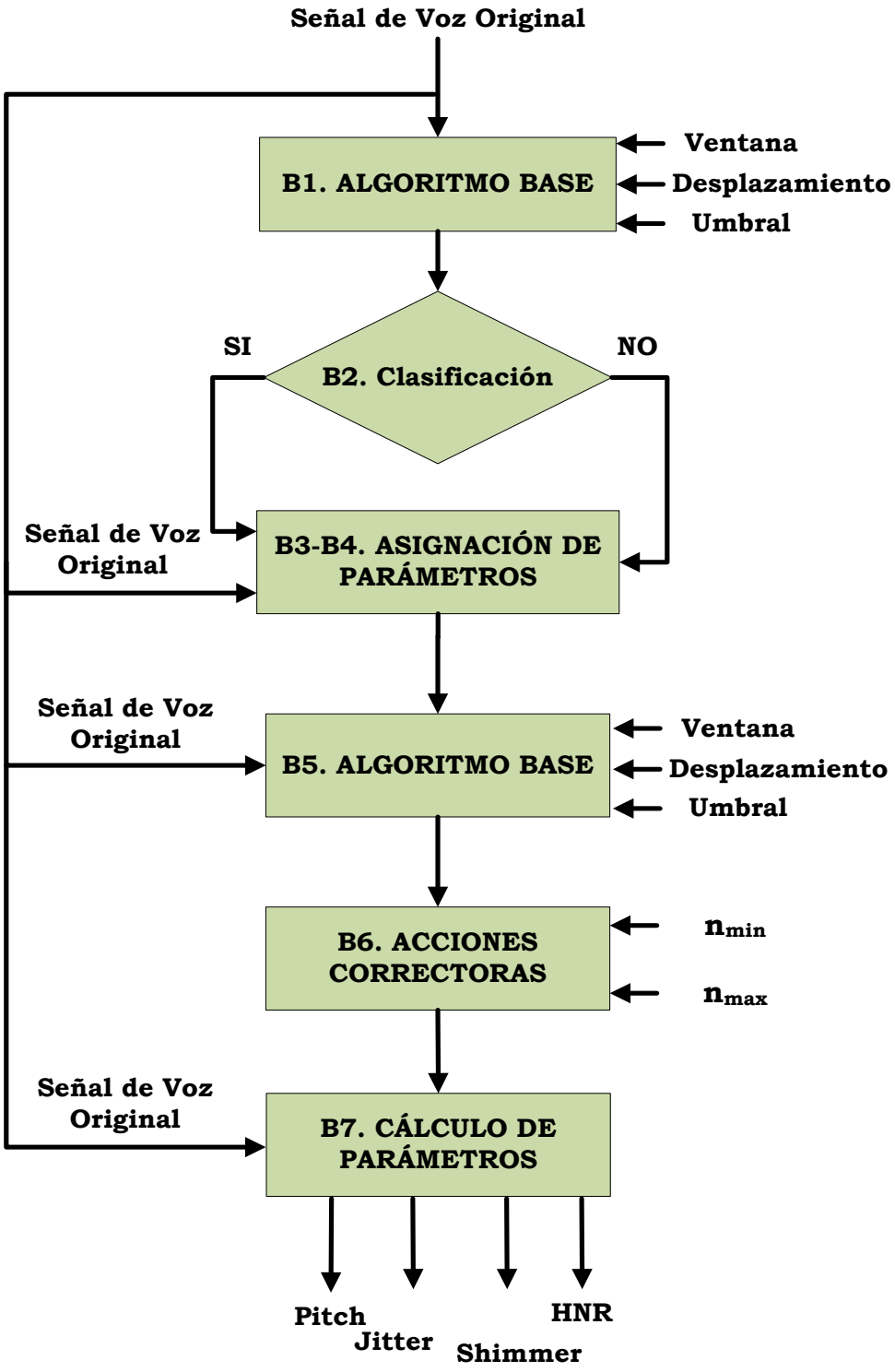

Figura 4.4: Esquema de "Mejora de la Parametrización de la Voz Esofágica"

A su vez, paralelamente a esta tesis hemos realizado paquetes de software que contienen algunos de los algoritmos diseñados en este trabajo de investigación. Estos paquetes de software van dirigidos al colectivo de los laringectomizados con objetivos de carácter social que también se hacen propios de esta tesis.

Estas mediciones nos han permitido ver las diferencias acústicas entre ambos tipos de voces. Es de resaltar que el Pitch o frecuencia fundamental en la voz esofágica es mucho menor que en la voz laringada, lo cual hace que la voz de las personas laringectomizadas sea más grave [Sano+89]. Estas diferencias son las que nos han guiado para elaborar los algoritmos que se describen a continuación.

Es decir, el algoritmo desarrollado que define con una mayor precisión los ciclos de la voz es tanto para aquellos casos en los que la voz está más dañada, como pueden ser la voz esofágica, como para los casos en los que la voz es sana o laringada. Por lo tanto, este algoritmo será válido para todo tipo de voces: sanas o laringadas y esofágicas. Una vez realizada una buena detección de los instantes de pitch o marcas, el cálculo del pitch es inmediato (Tabla 2.1) y se obtendrán los otros tres parámetros acústicos reconocidos por la comunidad científica como son: Jitter, Shimmer y HNR (Harmonic to Noise Ratio).

El cálculo automatizado de la posición de las marcas que definen los ciclos de voz y por ende, los parámetros acústicos, pasa por una *implementación ad-hoc* que cubra las carencias mostradas por las soluciones de los algoritmos de paquetes de software comerciales disponibles en la actualidad.

4.2 DISEÑO DE BAJO NIVEL

En este apartado se describen con detalle cada uno de los módulos que componen las dos etapas descritas anteriormente en el "**Diagrama de bloques del Algoritmo de la Voz Esofágica**" (Figura 4.1). En cada una de ellas, se detallan las transformaciones realizadas y el diseño de los algoritmos aplicados a las voces originales. Además de ello, se detallarán todos los sub-etapas de cada módulo indicando las entradas y salidas de los mismos.

4.2.1 Bloque de "Mejora de la Calidad de la Voz Esofágica"

En este apartado se describirán los distintos bloques del algoritmo de bajo nivel de la "**Mejora de la Calidad de la Voz Esofágica**" (Figura 4.1, Bloque A). A continuación se detalla el bloque A del mencionado algoritmo (Figura 4.5):

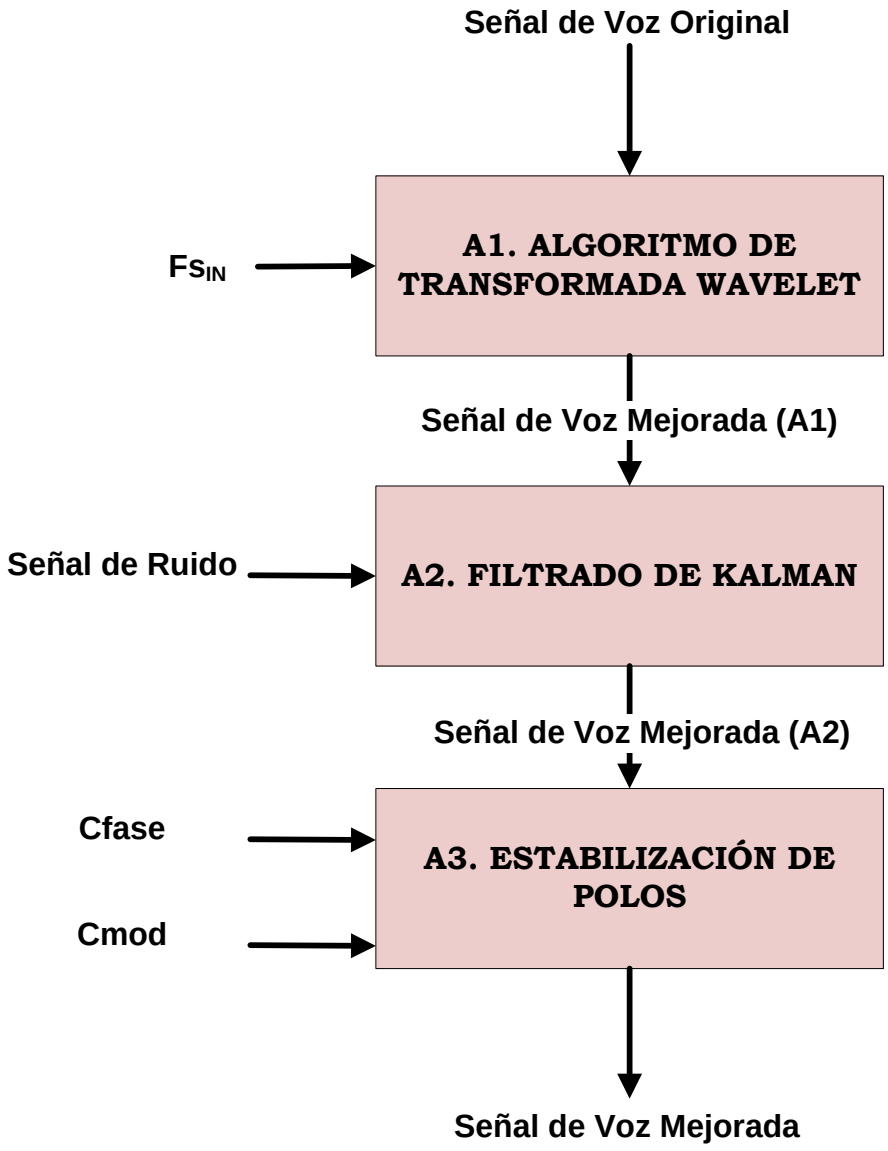

Figura 4.5: Diagrama detallado "Mejora de la Calidad de la Voz Esofágica"

En la etapa A1 se utiliza la técnica de procesado digital de señal de la transformada wavelet que está principalmente orientada a mejorar el shimmer y el tremor de la voz. Después, se propone una etapa en la que se mejorará el ruido

de la voz esofágica utilizando un filtrado de Kalman adaptado específicamente a las voces esofágicas (etapa A2).

Tanto la técnica de procesado digital de señal de la transformada wavelet como la de filtrado de Kalman se han diseñado propiamente teniendo en cuenta las especificidades y características de las voces esofágicas. Es decir, se han utilizado fragmentos de ruido de la voz esofágica, en periodos de silencio o donde no hay habla, y además, se han aprovechado las características de los parámetros de la voz esofágica en cuanto al pitch para mejorarla. Por último, se ha utilizado un algoritmo ya conocido como es la estabilización de polos para reduzca aún más el ruido de la voz esofágica (etapa A3).

Esta concatenación de algoritmos se presenta en un orden determinado. Se han realizado todas las permutaciones posibles de dichos algoritmos de procesado de señal observándose experimentalmente que este orden fijado es el que proporciona una mayor mejora la calidad de la voz esofágica.

Si nos fijamos en la tipología de una señal "tipo" de voz esofágica, como podemos ver en la Figura 4.6, vemos que tenemos una señal de voz con mucho ruido.

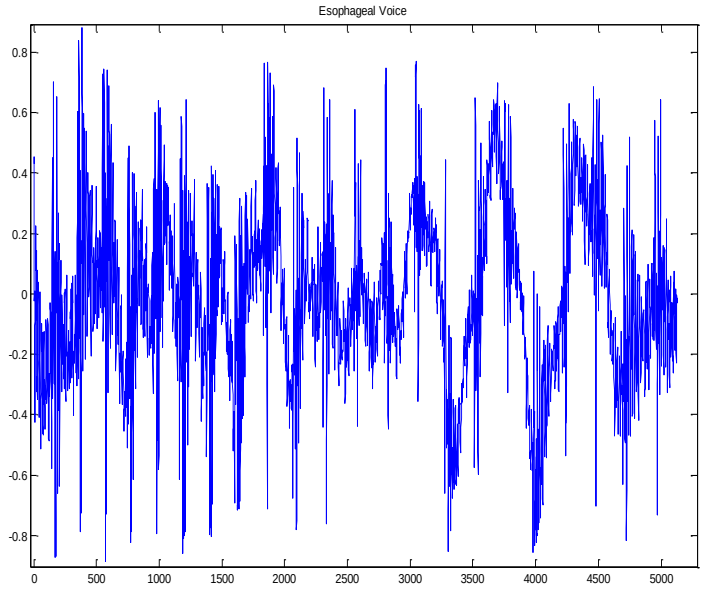

Figura 4.6: Señal de voz esofágica

Esta afirmación la podemos corroborar de forma empírica con la comparación con una señal de voz laringada o normal (ver Figura 4.7). Podemos observar que la voz esofágica tiene una gran variabilidad en la magnitud en los ciclos de voz o marcas de pitch (shimmer elevado), mientras que en la voz laringada es mucho menor. Esto significa que la señal de voz tiene un shimmer mayor que la voz normal. Además, se puede percibir un ruido de baja frecuencia en la voz esofágica que se llama tremor de la voz [Dromey+08].

Se puede observar en la señal de voz normal (ver Figura 4.7) que el ruido es mucho menor que en las voces esofágicas. Esto hace que el parámetro Harmonic to Noise Ratio (HNR) sea mucho menor que en las voces esofágicas (se puede distinguir en un menor color azul de la imagen de la Figura 4.7).

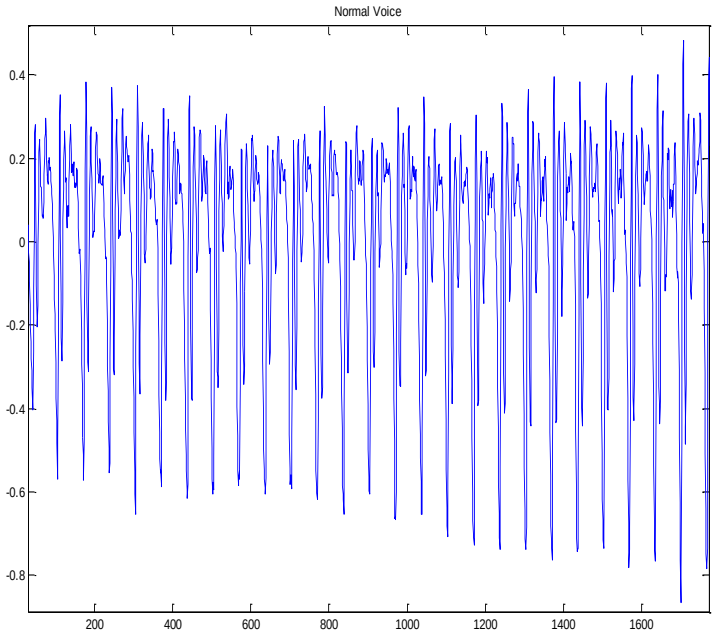

Figura 4.7: Señal de voz laringada o normal

Debido a la característica general de las voces esofágicas de un elevado ruido en la voz y un shimmer muy alto, se plantea algoritmos de procesado de señal que contrarresten estas dos carencias. Por lo tanto, se han planteado técnicas de procesado de señal que mejoren estos dos parámetros. Es por ello que la solución que se cree más adecuada es la presentada en la Figura 4.5.

4.2.1.1 Bloque del "Algoritmo de la Transformada Wavelet"

De cara a mejorar el ruido de baja frecuencia o el tremor de la señal y, a su vez el shimmer de la voz, se propone el bloque "Algoritmo de la Transformada Wavelet". El diagrama de bloques de este algoritmo se presenta en la Figura 4.8.

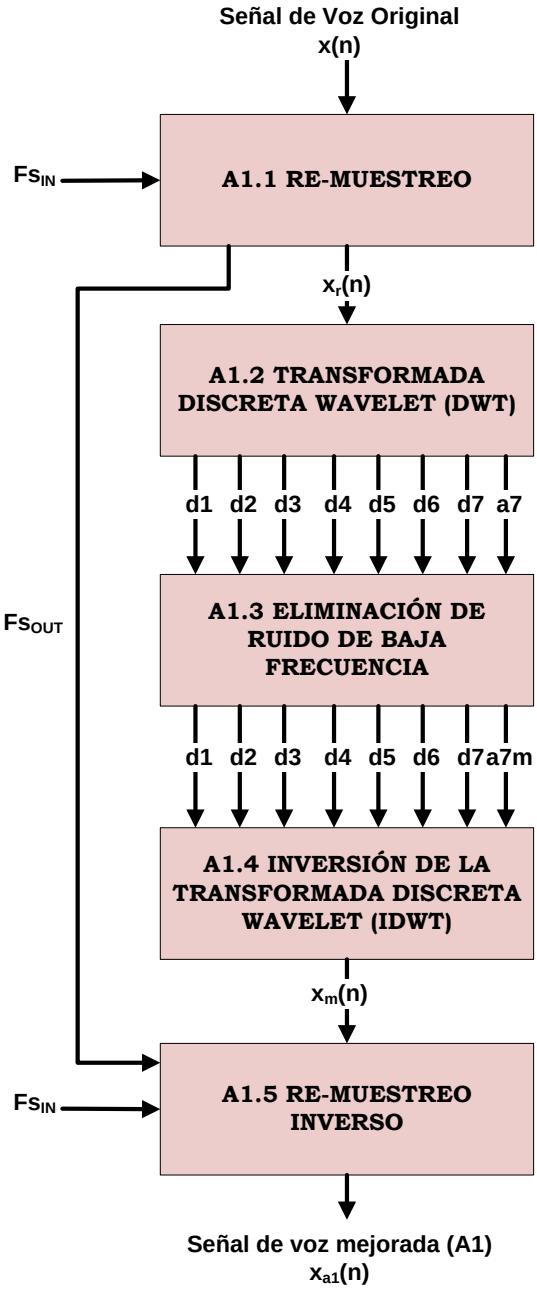

Figura 4.8: Diagrama de bloques del algoritmo de la trasformada wavelet

El algoritmo completo tiene como entrada o "**input**" una señal de voz original de la base de datos descrita en el capítulo 3 de esta tesis. La frecuencia de muestreo

de la señal, $F_{S_{IN}}$, se toma como parámetro para el re-muestreo de la señal y para el re-muestreo inverso de la señal de voz, es decir, para dejar la señal a la misma frecuencia con la que entra en el algoritmo. La señal después de la primera etapa queda re-muestreada y su frecuencia de muestreo es F_{SOUT}. Esta señal se re-muestrea para que después de aplicar la transformada wavelet las bandas de frecuencia queden en la misma banda del pitch. Esto se explica en el siguiente apartado, 4.2.1.1.1 Re-muestreo.

La salida o "**output**" del algoritmo es una señal de la voz mejorada (etiquetada como A1 por ser la salida de la etapa A1, ver Figura 4.5) con una frecuencia de muestreo, $F_{S_{IN}}$, igual a la señal de voz de la entrada.

4.2.1.1.1 Re-muestreo (A1.1)

Este primer bloque es un re-muestreo de la señal de voz. Tiene como entrada una señal de voz esofágica original etiquetada como x(n) y con una frecuencia de muestreo de $F_{s(in)}$ = 44,1 kHz y la salida es una señal re-muestreada a una frecuencia de muestreo de $F_{s(out)}$ = 12,8 kHz. La salida es la misma señal que la entrada pero re-muestreada y etiquetada como $x_r(n)$. La Figura 4.9 muestra las entradas y salidas de esta etapa.

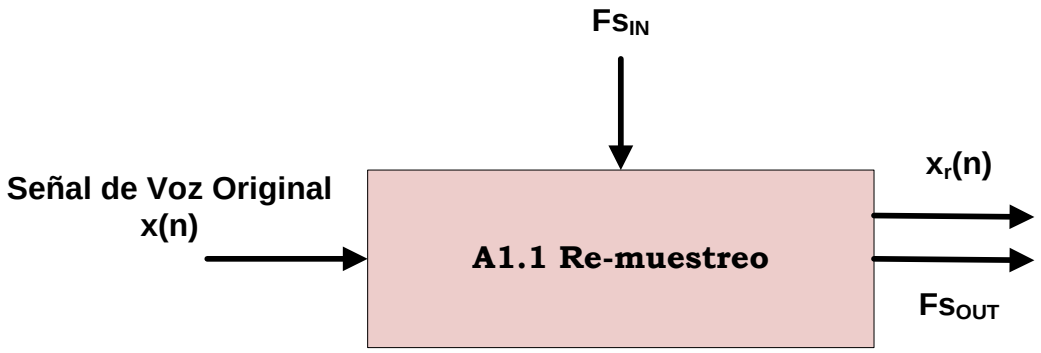

Figura 4.9: Bloque de pre-procesado (A1.1) del algoritmo de WT

La principal novedad de este algoritmo es que se escoge esta frecuencia de muestreo para que al realizar la transformada wavelet, las señales que se obtienen en las bandas de frecuencia de aproximación y detalle sean las adecuadas para la detección del pitch y poder así determinar las características

específicas de la voz esofágica. Además, se procesará la señal de bandas de frecuencia de entre 0 y 50 Hz dejando intacta la banda de frecuencia del pitch (50Hz – 100 Hz) para no distorsionar la señal. La técnica de subdivisión en bandas de frecuencia se ha utilizado para obtener el pitch [Kadambe+91], [Kadambe+92] [Nadeu+91] [Wing-kei+95].

4.2.1.1.2 Transformada Wavelet Discreta (A1.2)

Posteriormente al re-muestreo se realiza la transformada discreta wavelet (Discrete Wavelet Transform, DWT). Este supone el segundo bloque de la Figura 4.8.

Figura 4.10: Bloque de la Transformada Discreta Wavelet (A1.2)

Esta etapa tiene como entrada o "**input**" una señal de voz esofágica que ha sido re-muestreada y cuya frecuencia de muestreo es de $F_{s(out)}$ = 12,8 kHz.

La salidas o "**outputs**" son 8 señales de la misma longitud de la señal original. Debido a que se realiza la transformada wavelet discreta hasta el nivel 7, las señales de salida son 7 detalles ($d_1(n)$,..., $d_7(n)$) y la aproximación $a_7(n)$. La suma de estas señales produce la señal original.

El motivo de utilizar como wavelet madre "Bior 6.8" y 7 niveles se expone más adelante en este mismo apartado.

Como ya se ha mencionado en el apartado 2.2.2 de esta tesis, al realizar la transformada wavelet discreta a una señal, ésta se divide en dos señales dividiendo el espectro de la señal original. La aproximación es la parte del espectro de frecuencias más bajas y el detalle es la porción del espectro que está

en altas frecuencias. El mismo esquema de este bloque se puede observar con todos sus niveles en la Figura 4.11.

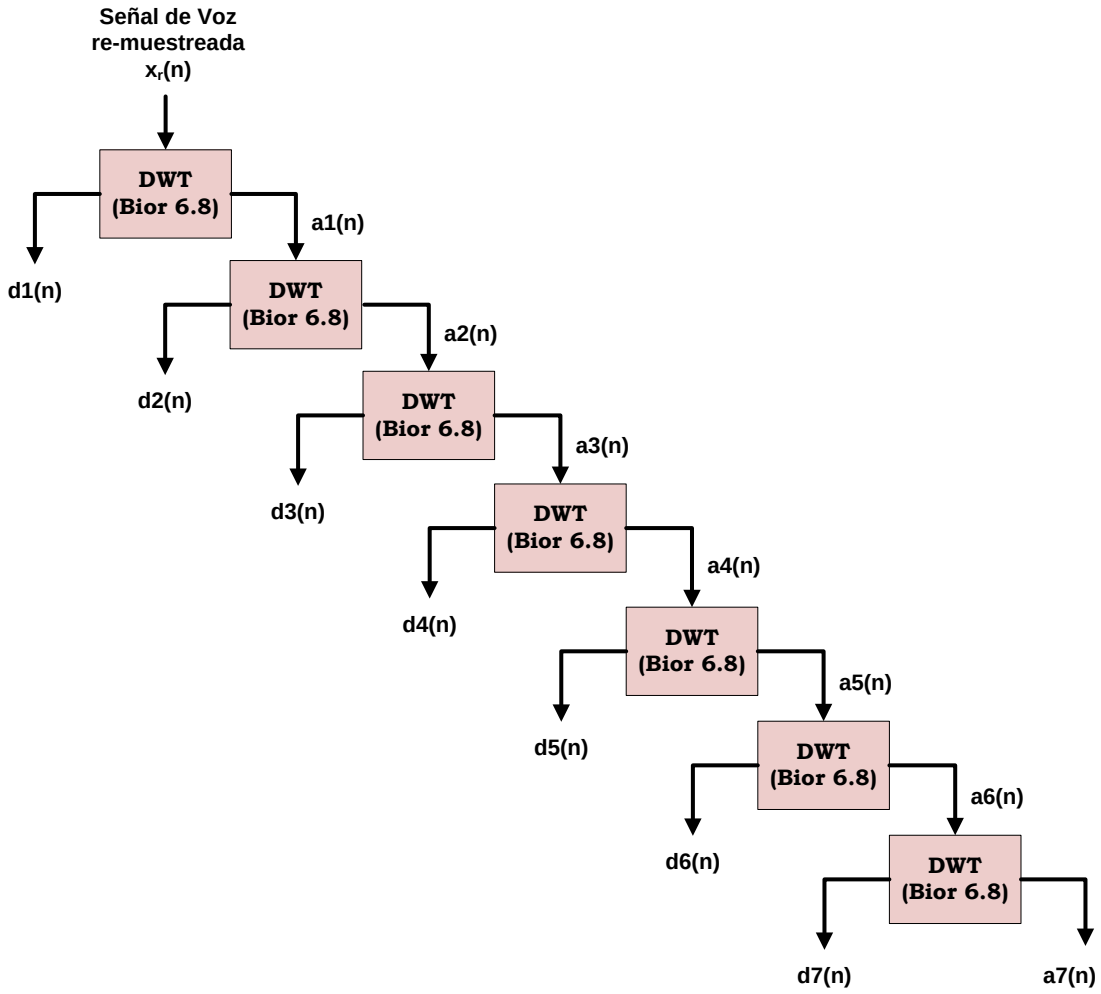

Figura 4.11: La Transformada Wavelet Discreta (DWT) en los 7 niveles

De esta manera, al realizar por primera vez la transformada discreta wavelet, la señal original re-muestreada a 12,8 kHz tiene como frecuencia máxima 6,4 kHz y se divide en dos señales, donde la señal en el ancho de banda de aproximación está entre las frecuencias 0 Hz y 3,2 kHz y, la señal en el ancho de banda de detalle está entre las frecuencias 3,2 kHz y 6,4 kHz. Obviamente, estos dos rangos de frecuencias no son los deseados donde se puede analizar el pitch de la voz, y tampoco se puede estudiar el ruido de baja frecuencia. Por lo tanto, se seguirá aplicando la transformada discreta wavelet (DWT) a la señal de aproximación, comprendida en frecuencias bajas, hasta obtener las señales en los rangos de

frecuencia deseados. Los rangos de frecuencia de cada señal al aplicar la transformada hasta el nivel 7 se puede observar en la Figura 4.12.

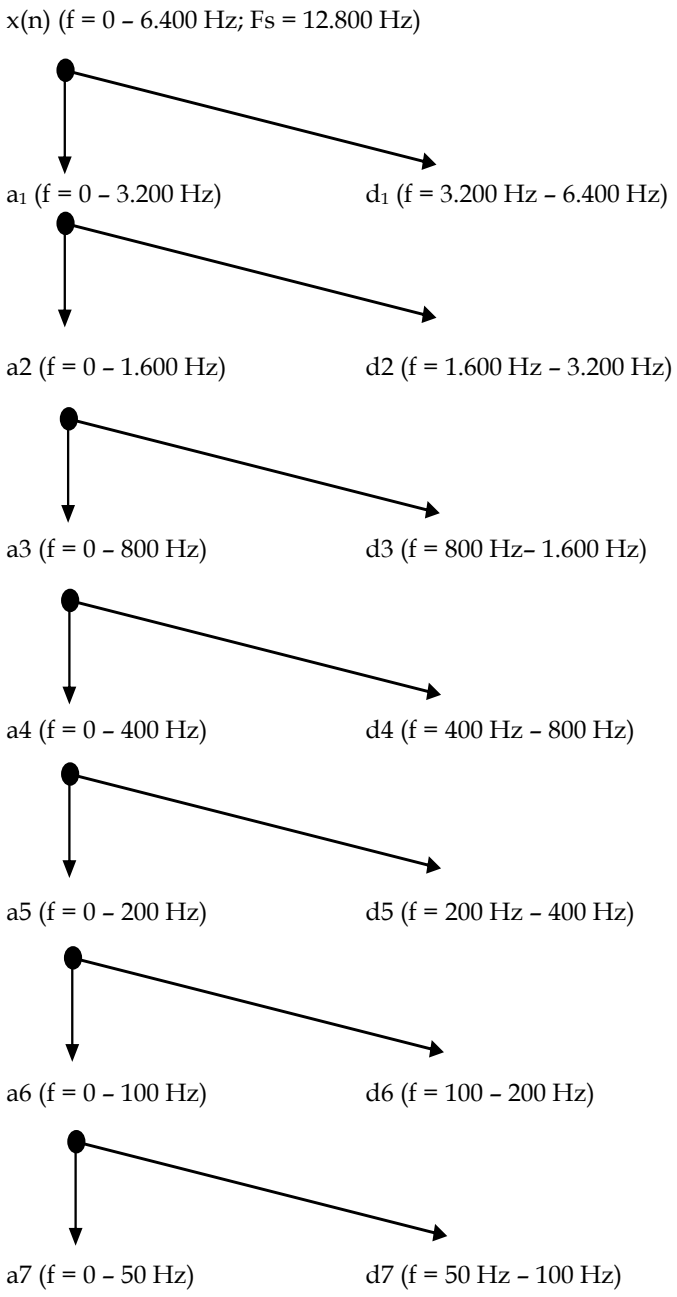

Figura 4.12: Rangos de frecuencia resultantes al aplicar las DWT

De esta manera, se obtienen las señales de los detalles de $d_1(n)$,..., $d_7(n)$ y la aproximación $a_7(n)$ tal y como se observa en la Figura 4.10. La suma de todas estas señales, $d_1(n)$,..., $d_7(n)$ y $a_7(n)$, da como resultado la señal que re-muestreada que entra en el bloque "Transformada discreta wavelet", es decir, la entrada ($x_r(n)$).

Cuando se realiza esta transformada es interesante comparar la señal original con las señales del último nivel. La principal novedad, como se puede ver en la Figura 4.13, es que en el último nivel, el nivel 7, quedan las señales deseadas: en el detalle $d_7(n)$, entre las frecuencias 50 Hz – 100 Hz, se encuentra la señal con la frecuencia fundamental o pitch. En la aproximación $a_7(n)$ tenemos el ruido de baja frecuencia de la señal de voz. A este ruido de baja frecuencia se le llama tremor de la voz [Dromey+08].

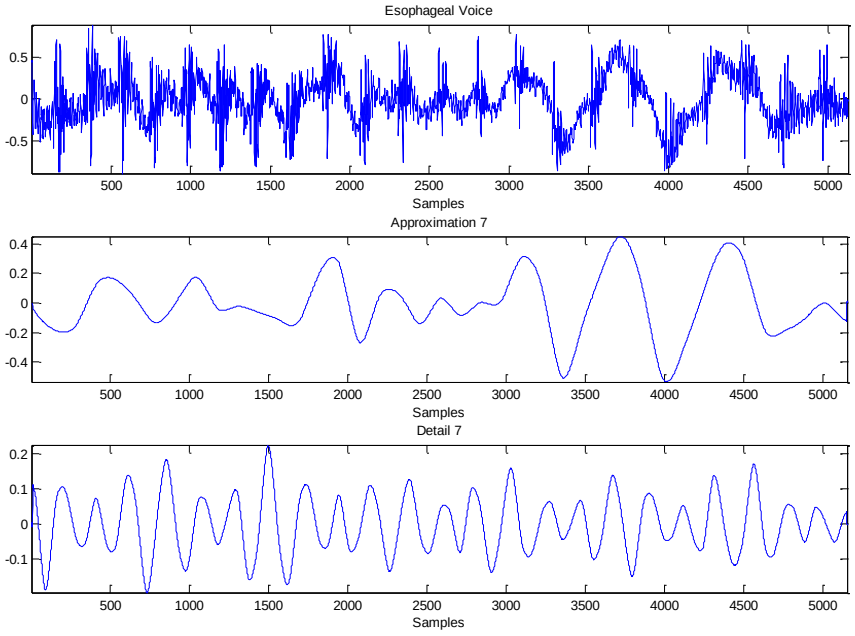

Figura 4.13: Señales después de aplicar la transformada wavelet discreta

Se puede observar en la Figura 4.13 cómo la aproximación en el nivel 7 ($a_7(n)$) es una especie de envolvente de la señal original. Éste es precisamente el tremor de la señal, ruido de baja frecuencia que ha de ser eliminado para la mejora de la calidad de la voz.

Como ya se ha comentado también en el apartado 2.2.2 existen numerosas familias de wavelet madre con diferentes características y órdenes [Tohidypour+10]. Algunas familias incluso tienen distintos órdenes para los filtros de reconstrucción y descomposición.

Tabla 4.1: Shimmer (dB) antes y después de aplicar el algoritmo

	Originales	Symlet	Bior 6.8	Coiflet	db6	Haar	Meyer
A1	0,594	0,51	0,106	0,53	0,372	0,0549	0,389
A2	0,468	0,363	0,227	0,524	0,369	0,565	0,374
A3	2,734	0,433	0,254	0,563	0,365	0,874	0,932
A4	0,573	0,831	0,372	0,751	0,809	0,919	0,823
A5	1,323	1,059	0,959	1,273	1,058	1,062	1,059
A6	0,796	0,677	0,521	0,642	0,624	0,73	0,819
A7	0,409	0,382	0,175	0,373	0,381	0,39	0,387
A8	0,342	0,391	0,186	0,384	0,391	0,38	0,383
A9	0,673	0,56	0,556	0,567	0,564	0,571	0,586
A10	0,339	0,537	0,171	0,676	0,528	0,669	0,629
A11	1,5	1,256	0,816	1,058	1,369	1,232	1,221
A12	0,412	0,491	0,289	0,484	0,415	0,492	0,474
A13	0,909	0,965	0,519	0,934	0,955	1,011	0,972
A14	0,868	0,564	0,346	0,56	0,55	0,587	0,573
A15	0,416	0,66	0,123	0,622	0,651	0,651	0,634
A16	0,23	0,603	0,264	0,654	0,582	0,694	0,704
A17	0,359	0,632	0,289	0,58	0,546	0,554	0,604
A18	0,71	0,803	0,413	0,772	0,823	0,861	0,746
A19	2,01	2,504	1,97	2,051	2,436	2,534	2,632
A20	0,343	0,538	0,124	0,501	0,925	0,483	0,493
A21	0,863	0,618	0,477	0,663	0,737	0,59	0,665
A22	0,997	1,318	0,272	0,272	0,287	0,335	0,398
A23	1,98	1,377	0,712	1,165	1,37	1,121	1,156
A24	0,494	0,759	0,275	0,519	0,794	0,874	0,533
A25	1,164	1,706	0,541	1,871	1,698	1,571	1,807
A26	2,069	0,812	0,599	1,543	0,813	0,755	0,808
A27	2,002	3,045	2,073	3,737	3,044	3,377	3,036
A28	2,133	2,361	1,151	1,048	1,151	1,177	1,141
A29	1,461	0,394	0,208	1,019	0,407	0,446	0,433
A30	2,772	1,956	1,566	1,989	1,657	1,702	1,636

Se han realizado muchas pruebas con numerosas wavelet madre y, finalmente, se ha comprobado que los mejores resultados son los que se obtienen con la wavelet biortogonal "bior 6.8". Es decir, el orden del filtro de reconstrucción es de 6 y el orden del filtro de descomposición es de 8. Estos resultados se pueden observar en la Tabla 4.1. Se ha aplicado el algoritmo de la transformada wavelet para diferentes wavelet madre. En concreto, se han utilizado las familias: Symlet, Biortogonales ("Bior 6.8"), Coiflet, Daubechies, Haar y Meyer. Estas familias wavelets son un amplio abanico de las ya existentes con lo que se puede apreciar las diferencias entre las distintas wavelets.

Se han medido todas las voces esofágicas (las 30 voces de la base de datos) antes y después del algoritmo. Se ha medido el shimmer en decibelios (dB) utilizando las diferentes familias wavelet madre. Observando la Tabla 4.1 se puede ver cómo el algoritmo funciona mejor cuando se utiliza "**Bior 6.8**" como wavelet madre.

Como se puede observar en dicha tabla, el algoritmo da mejores resultado utilizando "**Bior 6.8**" en todos los casos excepto en uno (señal A1). En los demás casos, los resultados con esta wavelet madre son mejores que las del resto de las wavelets madre. Los mejores resultados corresponden a las voces etiquetadas como A3, A5, A27, A29 y A30.

Para apreciar este hecho, se presenta la diferencia entre las diferentes wavelet madre con respecto al original (Figura 4.14). Es decir, la diferencia es igual a el shimmer (wavelet madre, columnas 2-7 de la Tabla 4.1) menos shimmer (voz original, columna 1 de la Tabla 4.1). Analizando la media de esta diferencia, se observa de forma clara que la menor media corresponde a la wavelet "Bior 6.8" con **-0,51 dB de media**.

Como se puede apreciar en la Figura 4.14, la línea roja, la wavelet madre "Bior 6.8", es la que tiene un menor valor con respecto a la original. Ésta es, por lo tanto, la mejor wavelet madre para este algoritmo ya que cuanto menor es el shimmer mayor es la calidad de la voz.

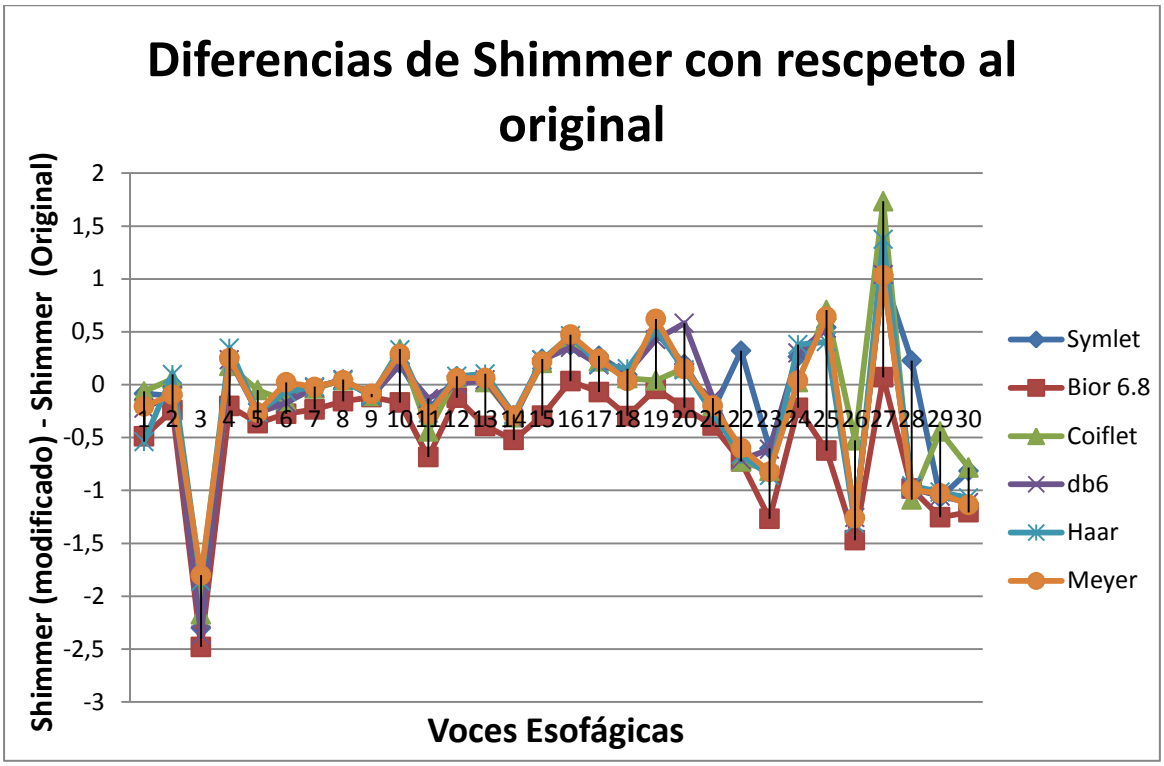

Figura 4.14: Comparativa del shimmer del algoritmo con respecto al original

Si analizamos los datos desde un punto de vista estadístico, se debe comentar que los datos del shimmer de las voces esofágicas originales no cumplen el criterio de normalidad, es decir, no podemos asumir normalidad en los datos. Esto determina en cierta manera el estudio estadístico a realizar. Si por una parte, el estudio más extendido es el "T-student" [Park+11], éste no es válido para un distribución de datos no normalizada. De cara a conocer la normalidad de los datos se han realizado dos pruebas: test de Kolmogorov-Smirnov [Fasano+87] y el de Shapiro-Wilk [Shapiro65]. Ambas pruebas han dado como resultado que la distribución de los datos del shimmer para la voz esofágica no está normalizada y la significancia de los datos es menor que 0,001 lo cual indica que no se pueden asumir la normalidad de los datos.

Una vez que no podemos asumir la normalidad de los datos, se realiza la prueba Wilcoxon [Wilcoxon45] para comparar los datos. Esta prueba nos muestra que los únicos datos que no son iguales son los originales y los datos tras realizar la etapa con la wavelet "Bior 6.8". La significancia de la prueba es menor que 0,001 ($p<0,0001$) y, por tanto, se rechaza la hipótesis nula que dice: "La mediana entre

los datos originales y los datos obtenidos tras la etapa wavelet con Bior 6.8 son iguales".

Haciendo la comparación entre los datos originales y los datos tras la etapa wavelet con otras wavelet madre vemos que no podemos inferir que los datos son diferentes. Las significancias con las diferentes wavelet madre son: Symlet (p<0,484); db6 (p<0,181); Haar (p<,139); Coiflet (p< 0,181); Meyer (p< 0,118). Ninguna de ellas es menor a 5% (p< 0,05) con lo que no se puede rechazar la hipótesis nula en la que dice que las medianas de los conjuntos de datos por parejas, tomando siempre una de ellas las originales, son iguales.

4.2.1.1.3 Eliminación de ruido de baja frecuencia (A1.3)

En esta etapa se elimina el ruido de baja frecuencia o tremor. En el bloque "Eliminación de ruido de baja frecuencia" tiene como entrada todas las señales mencionadas hasta ahora, es decir, $d_1(n)$,…, $d_7(n)$ y $a_7(n)$. La salida son las mismas señales excepto la $a_7(n)$ que ha sido modificada como se puede ver en la Figura 4.15.

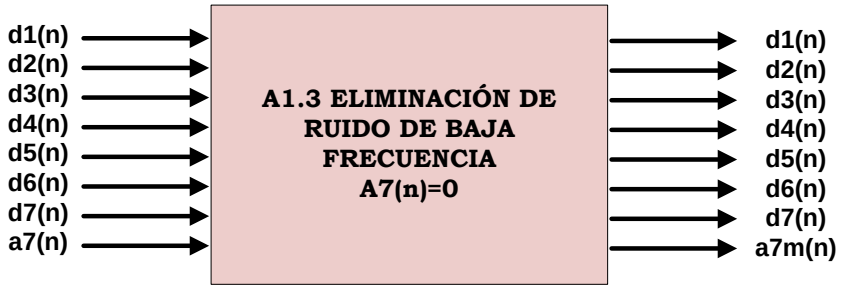

Figura 4.15: Bloque de eliminación de ruido (A1.3)

De todas señales de entrada, es la aproximación del nivel siete la que se modifica, eliminándola por completo. Es decir, se hace cero. Con este proceso se pretende atacar dos puntos importantes sobre la calidad de la voz: por una parte, al eliminar el ruido de baja frecuencia de la señal la amplitud de las marcas de pitch son más uniformes con lo que esto mejorará el shimmer y, por otro lado, se elimina el tremor de la señal original.

4.2.1.1.4 Inversión de la Transformada Wavelet Discreta (A1.4)

Una vez realizado este proceso, tenemos los detalles $d_1(n)$,…, $d_7(n)$ intactos y la aproximación $a_7(n)$ anulada. Con todas estas señales, en el bloque de "Inversión de la transformada discreta wavelet", se reconstruyen y se suman como se ha comentado anteriormente para obtener la señal reconstruida.

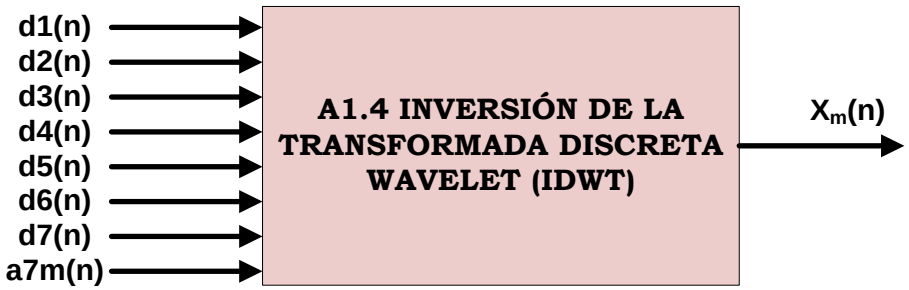

Figura 4.16: Bloque de inversión de la transformada wavelet (A1.4)

Las entradas o "**inputs**" de la etapa de inversión de la transformada wavelet son las señales $d_1(n)$,…, $d_7(n)$ y $a_7(n)$ procesada. La salida o "**output**" es una señal procesada que conforma la suma de todas las señales detalle y la señal de aproximación procesada como se puede ver en la Figura 4.17.

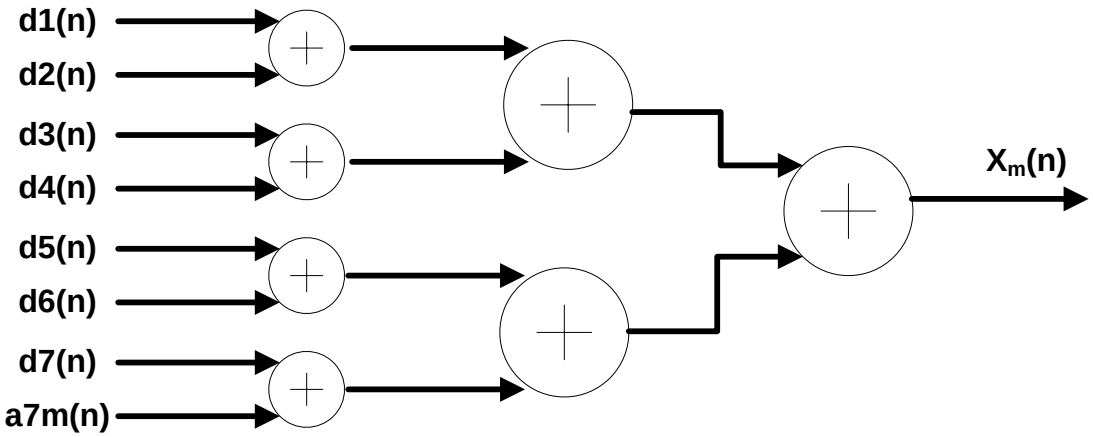

Figura 4.17: Reconstrucción de la señal procesada

Como se puede ver en la Figura 4.18 se ha eliminado el ruido de baja frecuencia y este hecho ha mejorado el shimmer de la señal y, por lo tanto, su inteligibilidad. Se puede comparar esta señal con la señal original mostrada en la Figura 4.13.

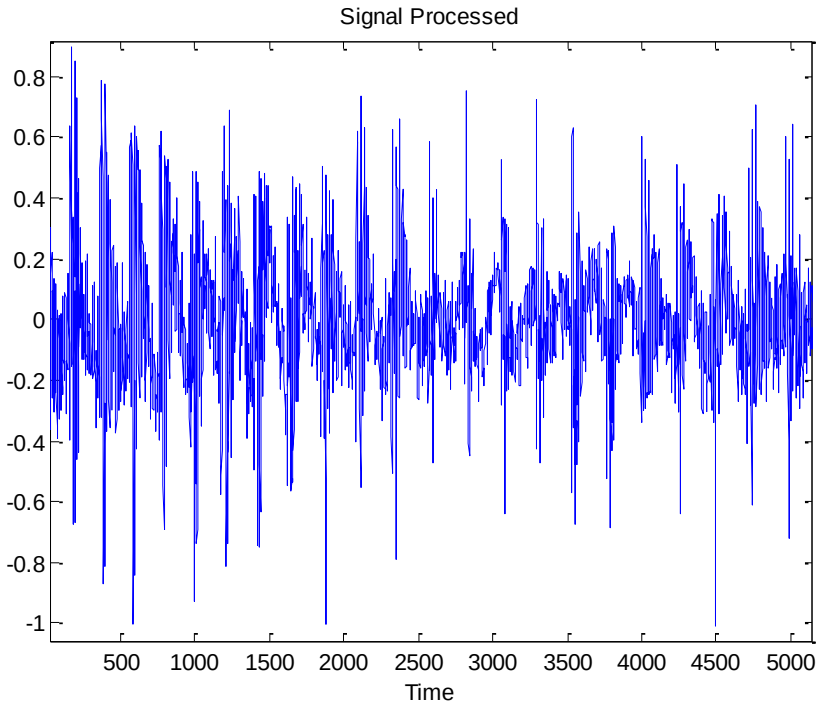

Figura 4.18: Señal de voz reconstruida tras el algoritmo de la WT

En esta señal se aprecia cómo el tremor o ruido de baja frecuencia ha desaparecido. Ahora la señal tiene una menor variedad en la amplitud de los picos de los instantes de pitch con lo que la calidad ha mejorado.

4.2.1.1.5 Re-muestreo inverso (A1.5)

Con todo el procesado de señal realizado, en el bloque "Re-muestreo inverso" la señal obtenida será re-muestreada a la frecuencia de muestreo original, es decir, a una frecuencia de Fs_{in}= 44,1 kHz.

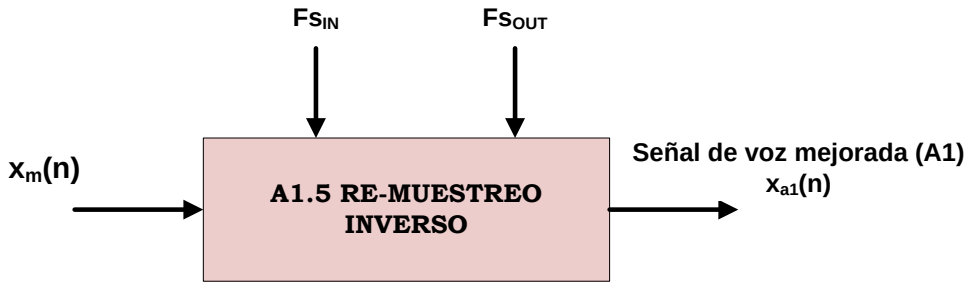

Figura 4.19: Bloque de post-procesado del algoritmo de la WT

La entrada o "**input**" de esta etapa es una señal de voz procesada con una frecuencia de muestreo de Fs_{out} = 12,8 kHz que es la señal de salida de la etapa de re-muestreo (A1.1). La salida o "**output**" es una señal re-muestreada de forma "inversa" a su frecuencia original, ya mejorada y con una frecuencia de muestreo de Fs_{in} = 44,1 kHz.

4.2.1.2 Bloque de "Filtrado de Kalman"

En este apartado se desarrollará el algoritmo del Filtro de Kalman, véase el segundo bloque Figura 4.5, específicamente aplicado a la mejora de la calidad de las voces de las voces esofágicas.

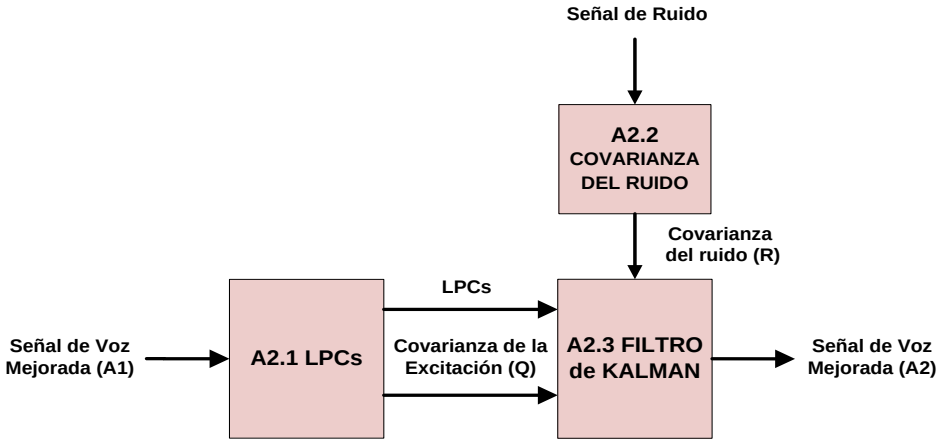

Figura 4.20: Bloque del algoritmo del filtrado de Kalman

La entrada de esta etapa es una señal de voz que previamente ha sido procesada por el bloque A1 de la Figura 4.5. Como se puede apreciar en dicha figura, la señal de entrada se ha etiquetado con el nombre "Señal de Voz Mejorada (A1)". Además de esta entrada proveniente de la etapa anterior, se utiliza una señal de ruido como ruido de medida del filtro de Kalman. El ruido utilizado en este algoritmo es ruido de aspiración de la voz esofágica. La salida de este bloque es una señal procesada con menor ruido y se ha etiquetado como "Señal de Voz Mejorada (A2)" (ver Figura 4.20).

La principal novedad de este algoritmo es que se ha adaptado el filtro de Kalman a las especificidades de las voces esofágicas. Es decir, se ha amoldado las

ecuaciones del sistema a la voz esofágica y además, se ha utilizado el ruido en los momentos de silencio como el ruido de medida. Este algoritmo mejora el elevado ruido que aún persiste después del bloque del "algoritmo de la transformada wavelet". Por lo tanto, este bloque incidirá en el parámetro de ruido de la voz, Harmonic to Noise Ratio (HNR).

En bloque el orden para el cálculo de LPCs que se ha aplicado es de 14, ya que se ha comprobado que da una precisión más que suficiente en el cálculo de los polos. También se podría hacer con un orden mayor que 14, pero esto no aportaría mayor información de relevancia y, sin embargo, provocaría un tiempo de procesado excesivo.

Como se ha mencionado en el apartado 2.2.3 de esta tesis, dados las observaciones pasadas y presentes, el filtrado de Kalman nos proporciona el estado estimado óptimo del método de los mínimos cuadrados.

La principal cuestión para desarrollar el algoritmo del filtrado de Kalman para la voz esofágica es cómo modelar el tracto vocal y el ruido de la señal de voz. Generalmente, la señal voz se estima por medio del modelo de autoregresión lineal (Autoregressive (AR) Model of Speech), [Gannot98] [Gibson+91] [Goh+99] [Paliwal+87].

4.2.1.2.1 Obtención de LPCs y Covarianza del ruido del sistema (A2.1)

La modelización que se utiliza en este caso es un sistema todo-polos con coeficientes predicción lineal, a_k, de orden p tal y como se ha explicado en el apartado 2.2.1 de esta tesis.

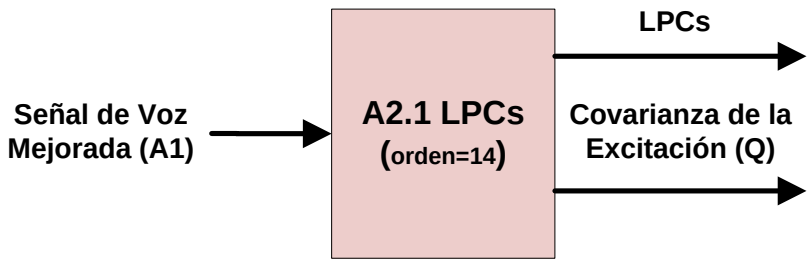

Figura 4.21: Bloque de obtención de los LPCs y la Covarianza del error (Q)

La entrada o "**input**" de este bloque es la señal que se ha procesado en el bloque A1 y que se ha etiquetado como "Señal de Voz Mejorada (A1)".

La salida o "**output**" de este bloque son los coeficientes de predicción lineal (LPC) y la covarianza del error o de excitación (*Q*). Esta covarianza es la que se utilizará posteriormente como la covarianza del *ruido del sistema, de planta o de proceso* en el filtrado de Kalman.

Por lo tanto, si denominados a la señal de voz sin ningún ruido *s(n)*, la modelización de coeficientes de predicción lineal (LPC) como combinación lineal de las *p* muestras anteriores es:

$$s(n) = \sum_{i=1}^{p} a_{si} s(n-i) + \omega(n) \qquad (4.1)$$

donde *s(n)* es la enésima muestra de la señal de voz y del ruido aditivo, respectivamente. Los coeficientes i-esimos del modelo autoregresivo (AR) de la señal de voz son a_{si}. La variable $\omega(n)$ es la enésima muestra de la variable aleatoria de error. Este error es una señal de ruido blanco estacionario, de media cero, y no correlacionado entre sí. El orden de los coeficientes de predicción lineal (LPC) utilizado para obtener la enésima muestra de la señal de voz y de ruido aditivo es $p = 14$.

4.2.1.2.2 Bloque de la covarianza de la señal de ruido (A2.2)

En este bloque se calcula la covarianza de la señal de ruido de entrada o ruido aditivo etiquetado como $v(n)$. La matriz de esta covarianza se calcula mediante la siguiente ecuación:

$$\mathrm{E}[v^T(k)v(j)] = R\,,\ k = j \qquad (4.2)$$

$$\mathrm{E}[v^T(k)v(j)] = 0\,,\ k \neq j \qquad (4.3)$$

Esta es la covarianza que se utilizará en el filtro de Kalman como covarianza del *ruido de medida*.

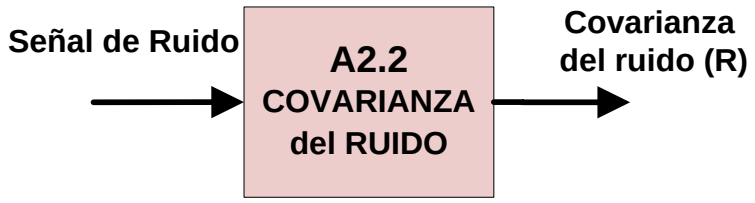

Figura 4.22: Bloque de la covarianza de la señal de ruido (A2.2)

La entrada o "**input**" de la señal es la señal de ruido utilizada para este algoritmo y la salida o "**output**" es la covarianza del ruido (*R*). En el siguiente apartado y una vez procesada mediante el filtro de Kalman se explicará qué ruido se ha utilizado en esta etapa.

4.2.1.2.3 Implementación del Filtro de Kalman (A2.3)

La última etapa de este bloque es la implementación del filtro de kalman con todas sus ecuaciones tal y como se presenta a continuación:

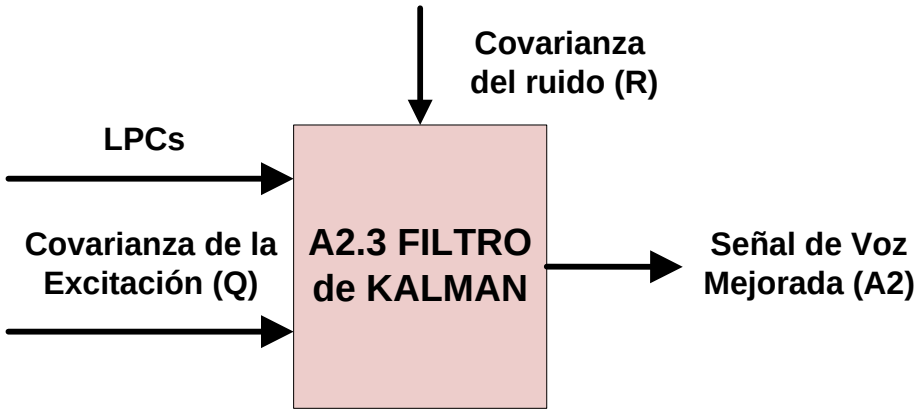

Figura 4.23: Bloque de implementación del filtro de Kalman (A2.3)

Las entradas o "**inputs**" de esta etapa son los coeficientes de predicción lineal (LPC) calculados en el bloque A2.1 y las covarianzas de los ruidos de planta o de sistema (*Q*) y del ruido de medida (*R*). Los coeficientes LPC se utilizarán para describir el sistema (A) y junto con las matrices *Q* y *R* se utilizaran en el filtro de Kalman. La salida o "**output**" será la señal de voz mejorada que etiquetaremos como A2. Se procede a continuación a describir el sistema. Si sumamos la ecuación 4.1 y la señal de ruido de entrada, es decir, la señal de voz y el ruido aditivo, obtendremos la señal de voz original:

$$y(n) = s(n) + v(n) \tag{4.4}$$

La siguiente cuestión para llevar a cabo el algoritmo de Kalman es que hay que expresar estas ecuaciones en el modelo de espacio de estados lineal invariante en el tiempo, tal y como se puede observar en la Figura 2.8.

La ecuación general del espacio de estados, ecuación 2.23, y teniendo en cuenta las ecuaciones anteriormente descritas, el diagrama de estados general se puede representar de la siguiente forma:

$$x(n) = A\,x(n-1) + G\,w(n) \tag{4.5}$$

$$y(n) = H\,x(n) + v(n) \tag{4.6}$$

donde la *A* es la matriz de estados y las variables aleatorias $w(n)$ y $v(n)$ representan el *ruido de planta* o *ruido proceso* y *ruido de medida*, respectivamente.

La matriz *A* puede tomar varias formas pero en este caso se ha optado por la forma canónica controlable. Los LPCs obtenidos en bloque A2.1 se introducen en la siguiente matriz y se representa de la siguiente manera:

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ a_p & a_{p-1} & a_{p-2} & \cdots & a_1 \end{bmatrix} \tag{4.7}$$

Las matrices H y G son matrices identidad de orden *p*:

$$H = G = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \tag{4.8}$$

Los vectores *x(n), y(n), v(n), w(n)* son las siguientes:

$$x(n) = [s(n-p+1), \ldots, s(n)]^T \tag{4.9}$$

$$y(n) = [y(n-p+1), \dots, y(n)]^T \quad (4.10)$$

$$v(n) = [v(n-p+1), \dots, v(n)]^T \quad (4.11)$$

$$w(n) = [w(n-p+1), \dots, w(n)]^T \quad (4.12)$$

En toda esta descripción del estado del sistema, las señales verdaderamente reales son la señal de voz, *s(n)*, y el ruido aditivo, *v(n)*, que se representa en el ruido de medida. El diagrama de bloques del modelo general del espacio de estados descrito en las ecuaciones 4.5 y 4.6 se puede observar en la Figura 4.24. Este diagrama es una variación del diagrama de estados general descrito en la Figura 2.8.

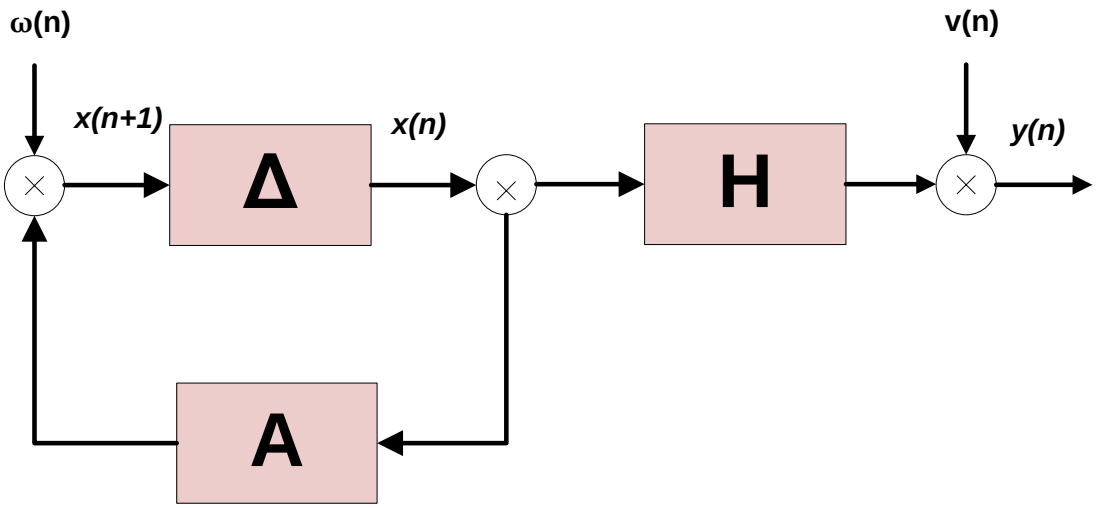

Figura 4.24: Diagrama de estados general adaptado a la voz y el ruido

El ruido aditivo o ruido de medida puede ser creado de forma aleatoria pero los resultados son mucho mejores si se utiliza el ruido coloreado. Se han realizado distintas pruebas con diferentes ruidos coloreados como son: ruido blanco, marrón, rosa, violeta y el ruido de la voz esofágica en los instantes de silencio. Es decir, ruido de instantes de silencio de la voz esofágica en las que el laringectomizado no está hablando pero mantiene un ruido de aspiración.

Para mostrar de forma adecuada estos algoritmos, se muestra a continuación una voz previa al procesamiento.

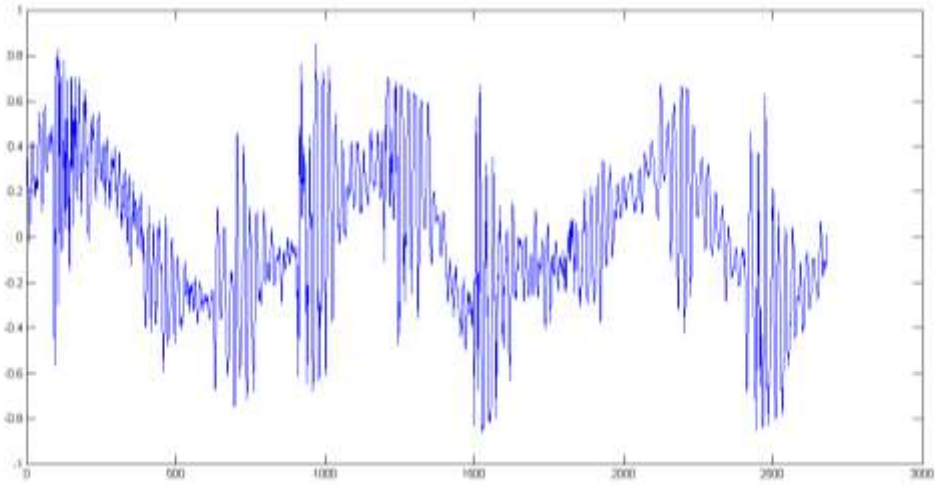

Figura 4.25: Voz previa al procesamiento

Esta misma voz después del procesamiento es la que se muestra a continuación:

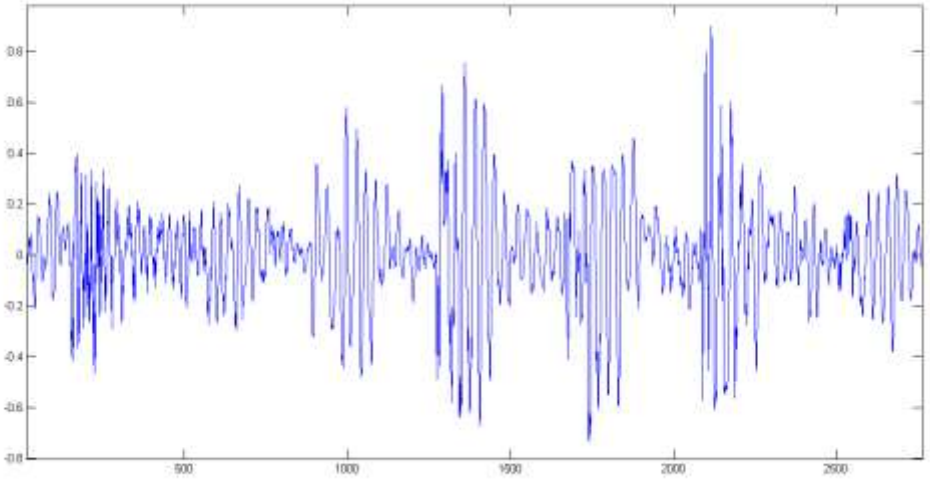

Figura 4.26: Voz después del procesamiento

Se puede observar que las magnitudes de la voz en los instantes de pitch y el ruido general de la voz es menor en esta última figura. Este menor ruido se puede apreciar tanto en las muestras alrededor de los instantes de pitch como entre periodo y periodo.

En la Tabla 4.2 se muestra los valores de HNR medidos en decibelios antes y después de aplicar el filtro de Kalman con los distintos ruidos mencionados. Se

puede observar que la voz procesada mediante kalman da un mejor resultado utilizando los instantes de silencio de la voz esofágica como ruido de medida. Esta es la principal novedad de este algoritmo, escoger el ruido de medida como el ruido en los instantes de silencio de la voz esofágica.

En dicha tabla, hay que destacar que las filas sombreadas son las que mejores resultados presentan. Concretamente, las voces etiquetadas como A3, A5, A11-A15 y A27 son los que mayor incremento presentan.

Tabla 4.2: HNR (dB) antes y después de Kalman con diferentes ruidos

	Originales	Blanco	Marrón	Voz esofágica	Rosa	Violeta
A1	-2,098	-0,452	0,409	2,037	0,288	-0,404
A2	-6,959	0,265	1,062	1,265	0,536	0,781
A3	-7,070	-6,185	-6,238	-4,954	-5,976	-5,805
A4	-7,979	-7,637	-6,678	-6,35	-6,721	-7,795
A5	-6,641	-5,836	-5,418	-4,393	-5,026	-7,978
A6	-0,802	3,658	3,205	3,938	2,591	2,648
A7	-9,191	-7,28	-6,779	-5,857	-6,067	-6,47
A8	-8,481	-6,783	-6,953	-5,202	-5,824	-7,013
A9	-2,526	-3,259	-2,026	-1,512	-1,856	-2,675
A10	-7,851	-4,589	-4,345	-3,057	-3,598	-5,237
A11	-7,298	-4,025	-5,954	-3,356	-4,561	-5,438
A12	-5,425	-4,176	-3,418	-2,782	-3,579	-3,929
A13	-5,700	-3,899	-3,659	-2,82	-3,762	-4,213
A14	-2,292	0,264	0,269	1,922	0,936	0,276
A15	-6,480	-5,462	-5,916	-4,349	-4,671	-4,627
A16	-5,687	-3,061	-3,733	-2,954	-3,498	-3,751
A17	-8,557	-4,851	-5,601	-3,525	-4,978	-5,131
A18	-7,735	-2,491	-2,045	-2,01	-2,802	-2,248
A19	-7,370	-5,976	-5,136	-4,862	-5,721	-6,657
A20	-6,570	-5,915	-6,754	-4,914	-5,614	-5,482
A21	-5,070	-2,887	-2,324	-1,645	-2,741	-4,053
A22	-5,040	-3,162	-2,871	-2,615	-2,936	-3,22
A23	-3,644	-1,951	-1,844	-1,012	-1,462	-1,486
A24	-5,010	-4,812	-4,802	-4,289	-4,524	-4,863
A25	-6,419	-5,852	-5,763	-5,096	-5,731	-5,591
A26	-9,309	-1,082	-1,466	-0,944	-1,618	-1,375
A27	-9,191	-6,504	-6,827	-5,781	-6,615	-6,106
A28	-6,638	-3,183	-5,045	-2,252	-4,051	-6,248
A29	-3,772	-2,615	-1,509	-1,427	-1,849	-2,674
A30	-7,796	-6,562	-5,672	-5,180	-5,561	-5,418

El incremento medio de las voces procesadas con el filtro de Kalman utilizando el ruido de los instantes de silencio es de 3,354 dB. La siguiente figura nos muestra la diferencia del HNR de las voces procesadas con el filtro de Kalman con respecto a las voces originales.

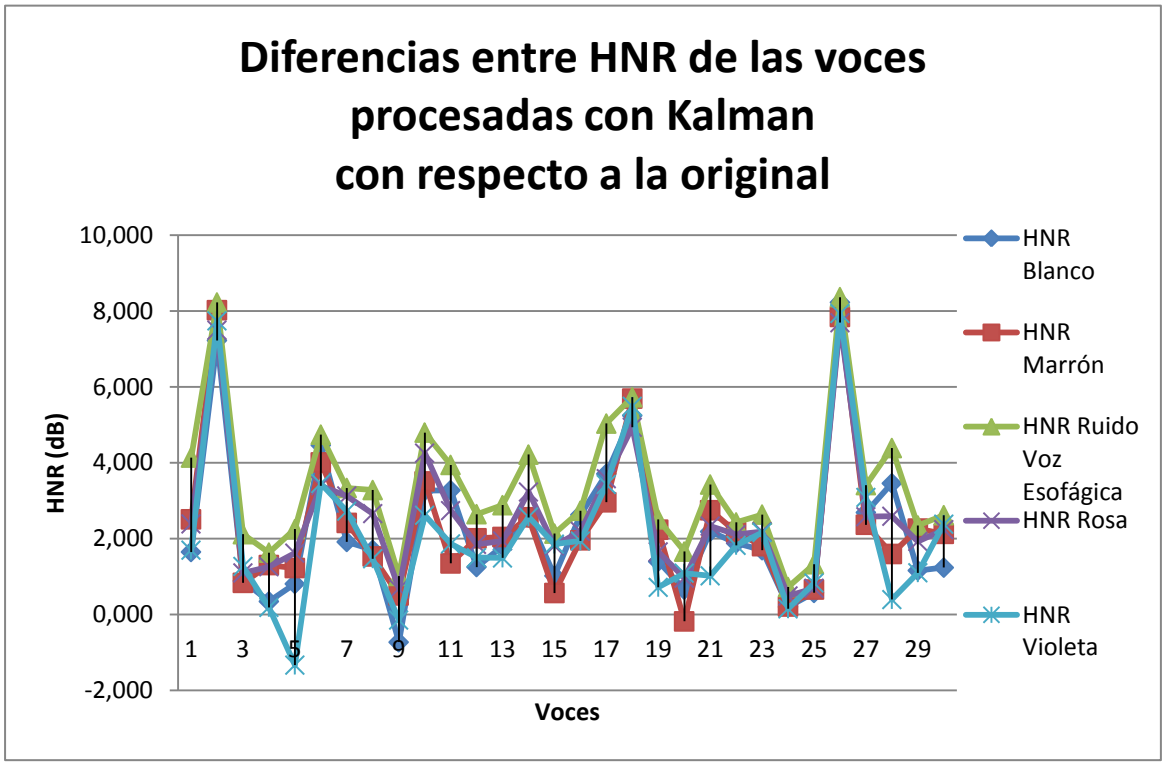

Figura 4.27: Comparativa del HNR del filtro de Kalman respecto al original

Si analizamos los datos desde un punto de vista estadístico, se debe comentar que los datos del HNR de las voces esofágicas originales cumplen el criterio de normalidad, es decir, podemos asumir la normalidad de los datos. Esto determina en cierta manera el estudio estadístico a realizar. En este caso, podemos utilizar la prueba estadística del "T-student" [Park+11]. De cara a conocer la normalidad de los datos se han realizado dos pruebas: test de Kolmogorov-Smirnov [Fasano+87] y el de Shapiro-Wilk [Shapiro65]. Ambas pruebas han dado como resultado que la distribución de los datos del HNR para la voz esofágica sí está normalizada y la significancia de los datos es mayor que 0,05 en todos los casos comparando todos los datos con el original, lo cual indica que se pueden asumir la normalidad de los datos.

Una vez que podemos asumir la normalidad de los datos, se realiza la prueba "T-student" [Park+11] para comparar los datos. Esta prueba nos muestra que los datos que no son iguales son: los originales y los datos tras realizar el procesado del filtro de Kalman con un ruido marrón, violeta y el ruido de silencio de la voz esofágica. Los ruidos blanco y rosa muestran una significancia mayor a un 5%, concretamente, p=0,142 y p=0,298, respectivamente.

Los ruidos que han mostrado una significancia menor al 5% arrojan los siguientes resultados: marrón (p=0,035); violeta (p=0,005) y ruido en los momentos de silencio de la voz esofágica (p<0,0001). Esto nos indica que con aunque todos ellos rechacen la hipótesis nula, en la que se dice que las medias de los datos cogidos por parejas es igual a la original; el que nos muestra mayor probabilidad de que los datos sean diferentes es el empleado con ruido en los instantes de silencio de la voz esofágica ya que presenta una probabilidad del 0,01% frente al 0,5% del ruido violeta y el 3,5% del ruido marrón. Esto junto con la gráfica presentada anteriormente nos muestra que es ruido esofágico el que nos da mejor resultado.

El ruido de proceso o de planta no suele ser deseado en los sistemas digitales. Aún así es un problema inevitable. En este caso, como la señal de voz es cuantificable, es decir, se puede expresar por medio de coeficientes LPC, el efecto del ruido de proceso puede ser minimizado sin un gran coste computacional. Como ya se ha mencionado anteriormente, con este ruido de planta dado por la ecuación 4.12 se obtiene la covarianza Q que se calcula mediante los LPCs de la voz esofágica.

Una vez que tenemos descrito el espacio de estados general adecuado al sistema de la voz esofágica, podemos aplicar el filtrado de Kalman. El diagrama de estados del filtro de Kalman es el que se muestra a continuación:

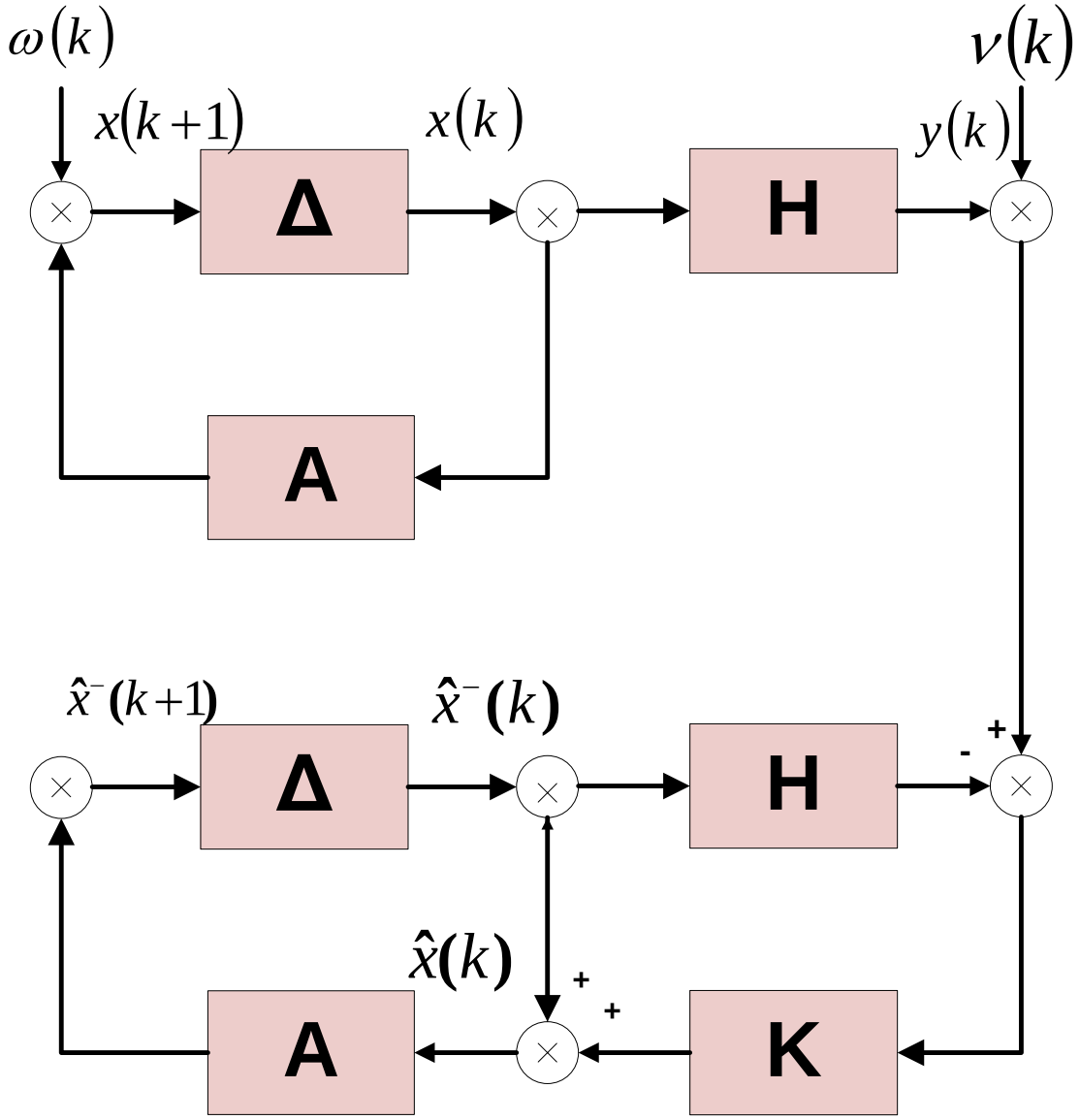

Figura 4.28: Diagrama de bloques del filtro de Kalman

De cara a enfatizar la corrección de la señal de la voz esofágica, el proceso del algoritmo de Kalman se realiza dos veces. La primera ejecución del filtro de Kalman mejora la señal pero los resultados se perfeccionan con la segunda pasada.

4.2.1.3 Bloque de "Estabilización de Polos"

El tercer bloque de la Figura 4.5 es el de "Estabilización de polos" y es el que presenta a continuación [García03]. Este algoritmo es el responsable de analizar y modificar los polos de la modelización del sistema del tracto vocal.

El algoritmo detallado de este bloque se ha presentado en el apartado 2.2.4 de esta tesis. En la Figura 4.29 se muestra el bloque general de "Estabilización de polos".

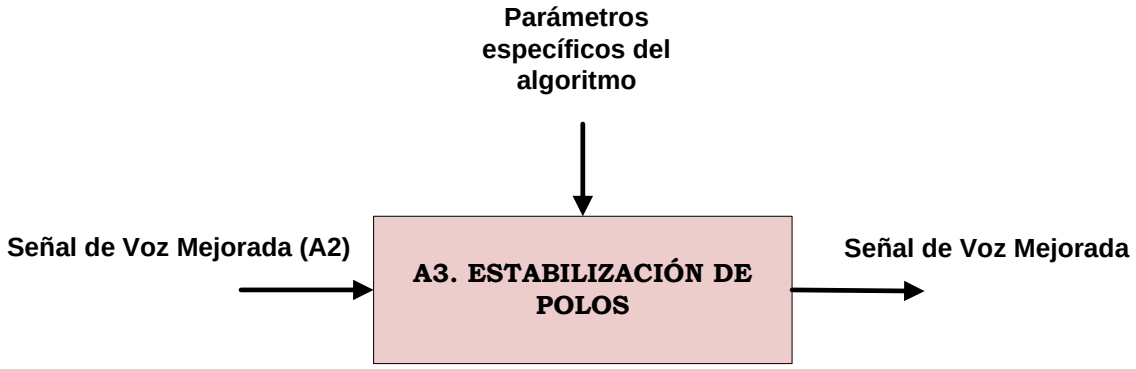

Figura 4.29: Bloque de "Estabilización de polos" (A3)

Este algoritmo tiene como entrada la señal mejorada proveniente de la salida del bloque del filtro de Kalman. Además, de esta entrada se introducen los parámetros específicos (Cmod y Cfase) explicados en el apartado 2.2.4 de la tesis. La salida de este bloque es la señal de voz mejorada con respecto al ruido.

4.2.2 Bloque "Mejora de la Parametrización de la Voz Esofágica"

En este apartado se describirán los distintos bloques del algoritmo de bajo nivel de la "Mejora de la Parametrización de la Voz Esofágica".

Cuando se trata de medir los principales parámetros de una señal de voz como el pitch, jitter, shimmer… se recurre generalmente a un paquete de software como el ya mencionado Multidimensional Voice Program (MDVP) [Deliyeski93]. Existen otros programas pero este es uno de los más populares entre la comunidad científica y uno de los más potentes. Al intentar medir cualquier parámetro con un paquete de software como el mencionado, dicho software establece los instantes de pitch de la voz con los que realizará sus cálculos y, como se puede apreciar en la Figura 4.30 b), si se realiza sobre una voz esofágica el resultado es erróneo.

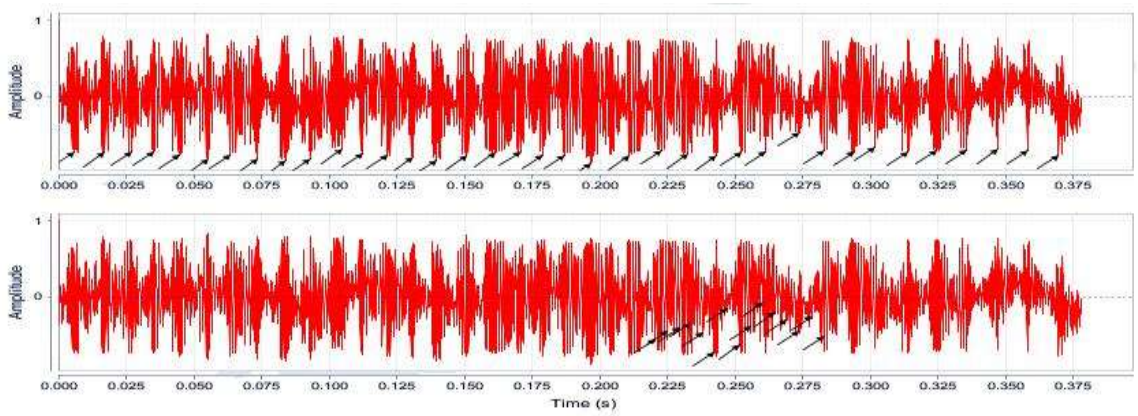

Figura 4.30: a) Instantes de pitch de una voz esofágica (arriba), b) Instantes de pitch de la misma voz obtenidas con el paquete de software MDVP (debajo)

Los instantes de pitch vienen marcados con flechas negras sobre la medición de la misma voz esofágica (ver Figura 4.30). En la imagen superior vemos los instantes de pitch correctamente colocados sobre la señal de voz esofágica y que han sido calculados con el algoritmo que a continuación se describe. En la imagen inferior, se muestra la extrapolación de la figura obtenida con el MDVP de la misma señal de voz esofágica y con las marcas de pitch absolutamente erróneas.

Con este ejemplo queda evidenciado que los paquetes de software comerciales junto con los algoritmos que utilizan, en general, no son propicios para analizar voces esofágicas. Por ello, se propone un algoritmo que de forma automatizada sea capaz de establecer los instantes de pitch de forma correcta para cualquier tipo de voz: las voces sanas y las voces esofágicas.

La entrada o "**input**" del algoritmo es una señal de voz, que como se ha comentado, puede ser una voz tanto sana como esofágica. Para el correcto desarrollo del algoritmo es necesario establecer ciertos parámetros que se comentarán con mayor detalle en la explicación de cada una de las etapas. La salida o "**output**" del algoritmo es el pitch corregido con un vector con los instantes de pitch de la señal de voz. La Figura 4.31 muestra el diagrama de bajo nivel de "Mejora de la Parametrización de la Voz Esofágica.

El algoritmo está basado en un procedimiento iterativo de obtención de los picos negativos de la voz. Se utilizan los picos negativos ya que se ha observado que los picos negativos son los que más energía tienen. El algoritmo se basa en una función que extrae los instantes de pitch o las marcas de la señal de voz para calcular cada componente de pitch.

Las iteraciones se realizan para una mayor precisión del algoritmo, es decir, para obtener una mejor medida de la voz dependiendo del tipo de voz que se esté analizando. Dependiendo de cada tipo de voz, dicha señal tendrá un rango de pitch más probable con lo que se ajustarán debidamente los parámetros del algoritmo [Karthikeyan+05]. Por lo tanto, el algoritmo en una primera etapa realiza una obtención del pitch de forma preliminar y orientativa.

Dependiendo del valor obtenido en la fase anterior se realiza, a continuación, una clasificación de la voz teniendo en cuenta dos tipos de voces: sanas o esofágicas. Posteriormente, se llevan a cabo unas acciones correctoras para eliminar o detectar, según el caso, instantes de pitch no obtenidos en las etapas anteriores. Una vez realizado todo este proceso se obtiene el pitch corregido para cada una de las voces esofágicas.

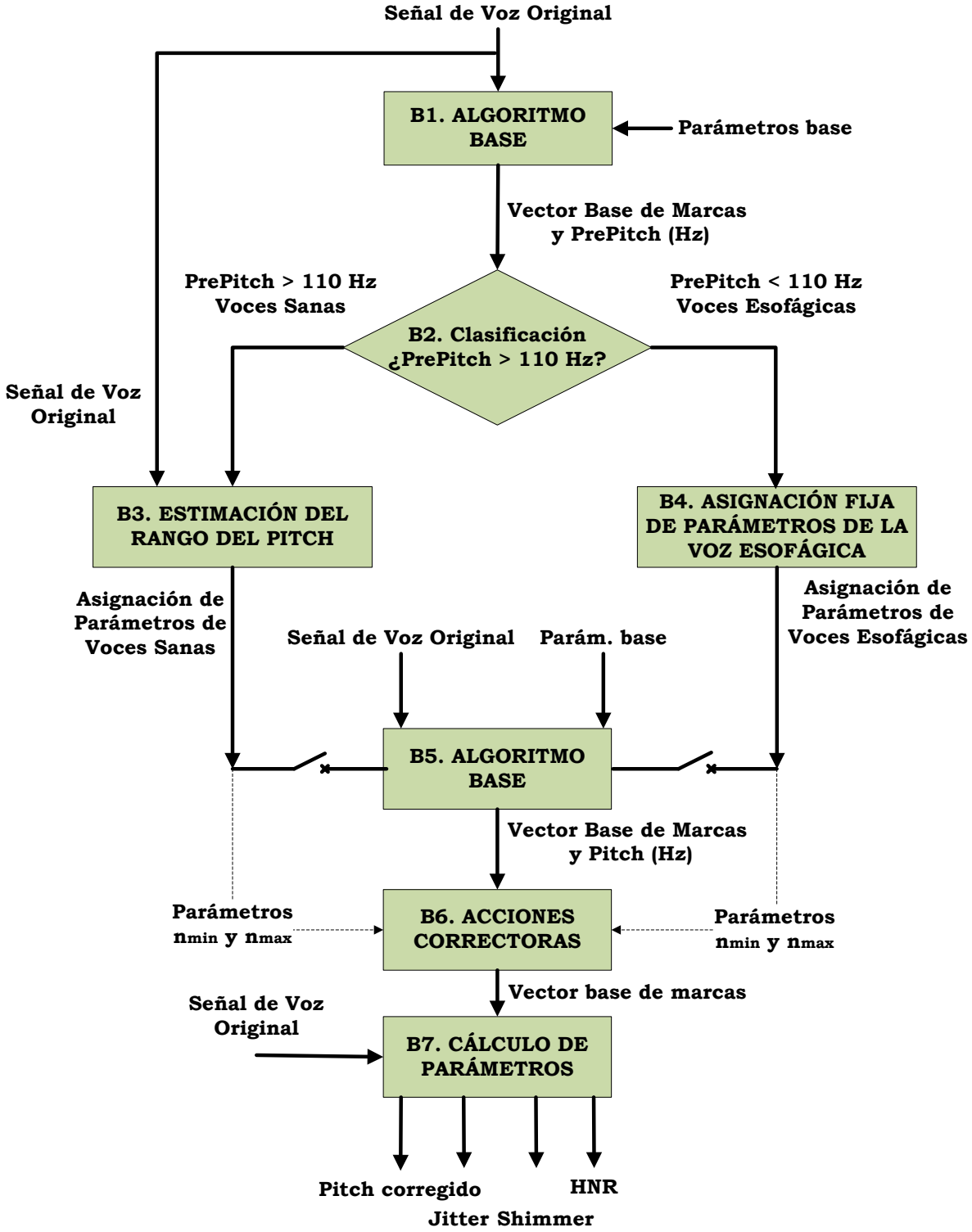

Figura 4.31: Diagrama de bajo nivel de la parametrización de la voz

4.2.2.1 Bloque "Algoritmo Base"

El "**Algoritmo Base**" de la Figura 4.31, en el bloque etiquetado como B1, ha sido diseñado para la correcta determinación de los instantes de pitch o ciclos de la voz.

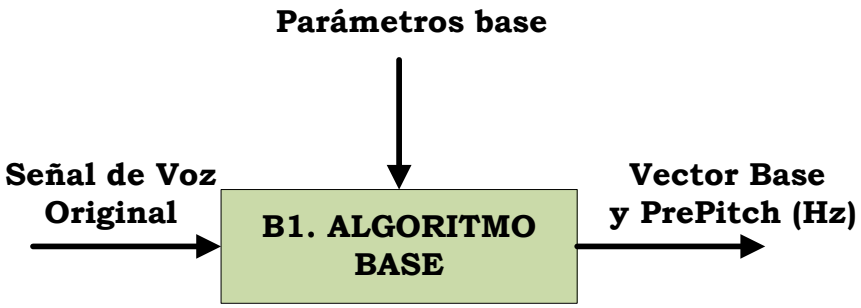

Figura 4.32: Algoritmo Base (B1)

La entrada o "**input**" de esta etapa es la señal de voz original. En este caso como ya se ha comentado dicha señal puede ser sana o esofágica. Para el correcto desarrollo del algoritmo es necesario introducir unos parámetros base que en esta primera etapa son: el tamaño de la ventana, el tamaño del desplazamiento de la ventana (ambas dos medidas en número de muestras) y un parámetro que indicará el umbral de la magnitud de los picos de la señal. Concretamente, en esta etapa se establece empíricamente con 64 muestras para el tamaño de la ventana, 10 muestras para el tamaño del desplazamiento de la ventana y 0,2 como factor de la máxima amplitud para establecer el umbral de los picos de la señal. Es decir, se establece como un umbral el 20% del máximo absoluto de la señal de voz.

La salida o "**output**" es un vector denominado base ya que es el que se toma como referencia para obtener el pitch con todo ceros excepto en las posiciones donde no existe un mínimo relativo y la magnitud en la posición donde aparece dicho mínimo. Una vez obtenido este vector con la magnitud de los instantes de pitch, es inmediato obtener el valor del pitch utilizando las ecuaciones de la tabla 2.1 del apartado 2.3.1 de esta tesis.

El algoritmo detallado del "Algoritmo Base" se muestra en la Figura 4.33.

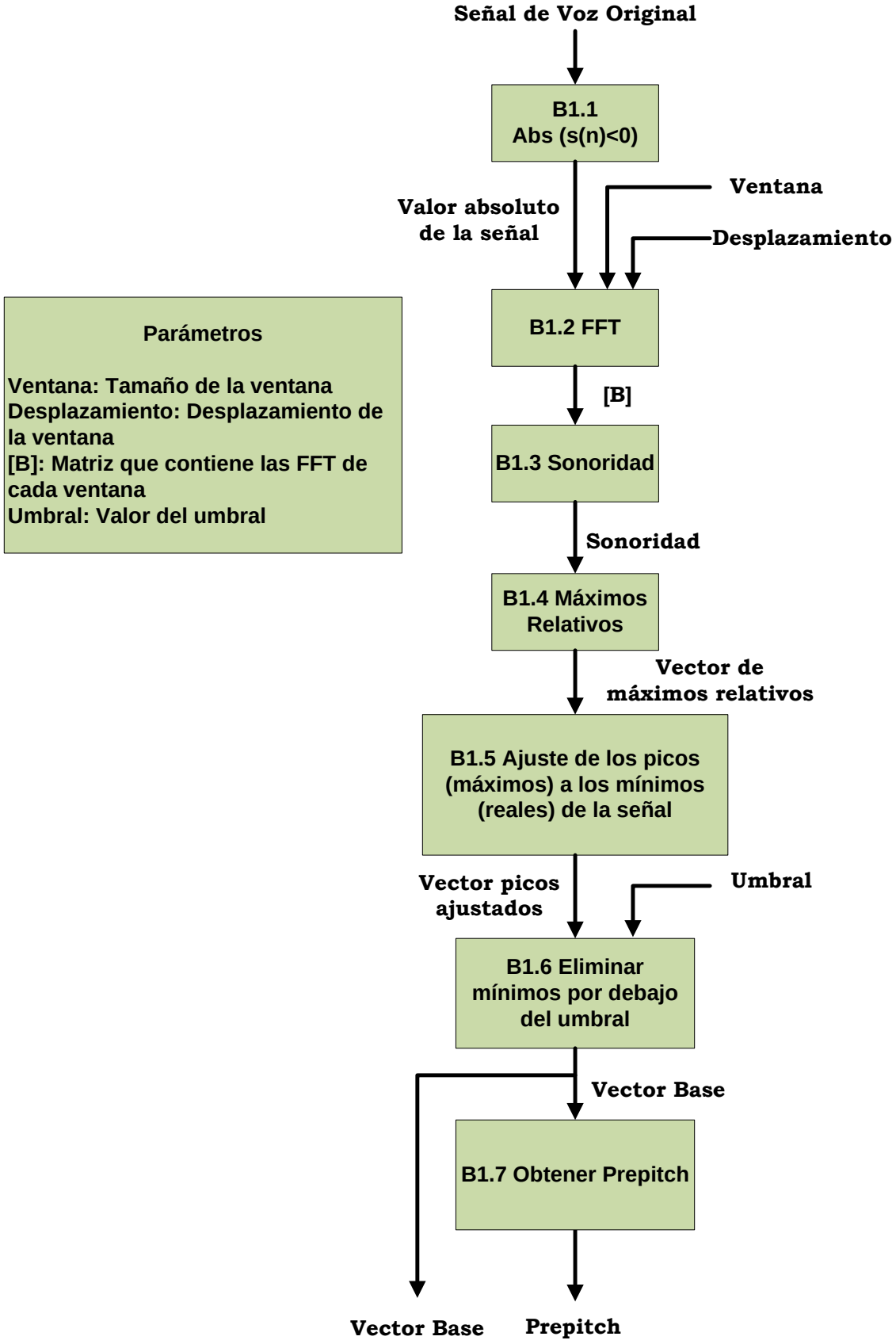

Figura 4.33: Organigrama detallado del Algoritmo de Base

4.2.2.1.1 Obtener el valor absoluto de la señal de entrada (B1.1)

El organigrama de la Figura 4.33 nos muestra la etapa completa del "**Algoritmo Base**" de manera más detallada. En él se observa que en la primera etapa se realiza el valor absoluto de la señal.

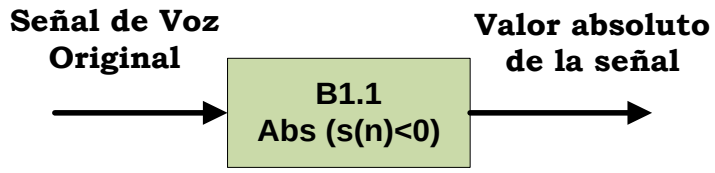

Figura 4.34: Bloque de valor absoluto de la señal (B1.1)

La entrada o "**input**" de esta etapa es la señal de voz original que como ya se ha comentado es una señal de la base de datos que puede ser sana o esofágica. La salida o "**output**" es el valor absoluto de la señal original (ver Figura 4.34).

Se realiza este proceso porque en la mayoría de las voces los picos con más energía son los negativos y realizar el valor absoluto de la parte negativa facilita la detección de los máximos (o mínimos, según el caso) relativos.

4.2.2.1.2 Fast Fourier Transform (B1.2)

Una vez realizado el valor absoluto de la señal, se realiza la transformada rápida de Fourier (Fast Fourier Transform, FFT) [Fourier22].

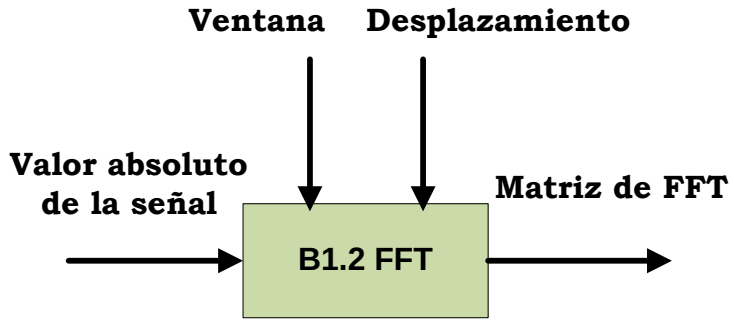

Figura 4.35: Bloque de la Transformada Rápida de Fourier (FFT) (B1.2)

Para realizar esta etapa se toma la entrada o "**input**" la salida de la etapa anterior, el valor absoluto de la señal, y se divide en ventanas o tramas (tamaño de 64 muestras para este bloque y varía en el bloque B.4) con un cierto desplazamiento

(tamaño de 10 muestras y varía en el próximo bloque de algoritmo base, B.4). Estos valores son una asignación de parámetros que se realiza para el correcto funcionamiento del algoritmo. Con este proceso se obtiene como salida o "**output**" una matriz con las transformadas de Fourier de cada una de las tramas para posteriormente obtener la sonoridad (ver Figura 4.35).

4.2.2.1.3 Sonoridad (B1.3)

Una de las novedades del algoritmo desarrollado es que está basado en la medida de la **sonoridad**. La sonoridad se define como sigue:

$$Sonoridad = \sum_{n=1}^{N} |B(k,n)| \qquad (4.13)$$

siendo B(k,n) una matriz que contiene las Transformadas Rápida de Fourier (Fast Fourier Transform, FFT) de N puntos de las k ventanas en las que se halla dividido la señal. Esta es, precisamente, la entrada de la etapa "Sonoridad". La salida es un vector con el cálculo de la sonoridad de las distintas tramas (ver Figura 4.36). Se puede observar en la figura que este bloque tiene dos pequeñas etapas: una es el cálculo de sonoridad y la otra es un filtrado de ésta.

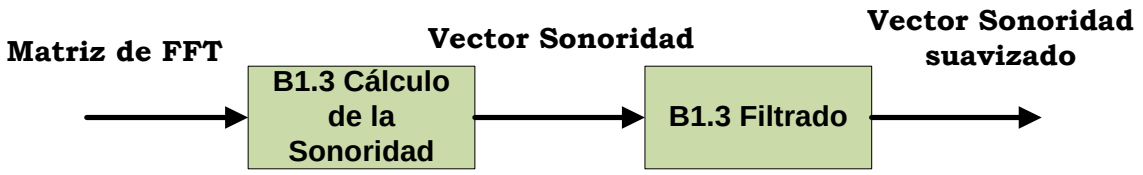

Figura 4.36: Bloque de la Sonoridad (B1.3)

La sonoridad se pasa por un filtro cuya función de transferencia en el dominio Z es:

$$H(z) = \frac{1+z^{-1}+z^{-2}+z^{-3}+z^{-4}}{5} \qquad (4.14)$$

El motivo de este filtro es simplemente de suavizado de la función.

De la ecuación presentada se deduce que la sonoridad es la *energía contenida en una ventana* pudiendo establecerse la hipótesis de que a mayor sonoridad en un tramo de señal, mayor es la probabilidad de encontrar una marca en el mismo.

Se puede ver cómo se desarrolla la hipótesis de partida, buscando los máximos relativos de la curva de sonoridad para identificar los mínimos absolutos en cada tramo, eliminando en última instancia aquellos por encima de cierto umbral.

4.2.2.1.4 Máximos relativos (B1.4)

Una vez que se ha calculado la sonoridad y se ha suavizado, se obtienen los valores de los máximos relativos. Aquí es donde se ubicarán los mínimos relativos de la señal que proporcionan los instantes de pitch.

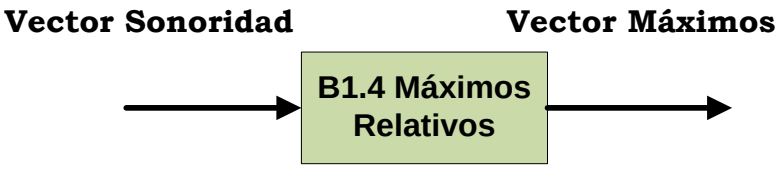

Figura 4.37: Bloque de Máximos relativos (B1.4)

La entrada de esta etapa es un vector que contiene los valores de sonoridad de cada una de las tramas de la señal. La salida es un vector con los máximos relativos del vector sonoridad (Figura 4.37). Es decir, en las posiciones del vector donde haya un máximo relativo aparecerá su magnitud y en las demás posiciones habrá ceros.

4.2.2.1.5 Ajuste de los picos (máximos) a los mínimos (reales) de la señal (B1.5)

La longitud y la amplitud del vector de máximos son parámetros que se deben ajustar para que encajen con los valores de la señal original. Para ello, el primer paso es invertir la señal de cara a obtener los máximos relativos en la parte negativa de la señal de voz, es decir, los mínimos. El próximo paso es ajustar cada mínimo a su posición real en la señal original. Para ello se toma la trama de la señal original, después se centra en la posición de cada mínimo obtenido previamente y, el vector resultante se construye insertando el valor absoluto de cada mínimo en su posición y el resto se deja a cero.

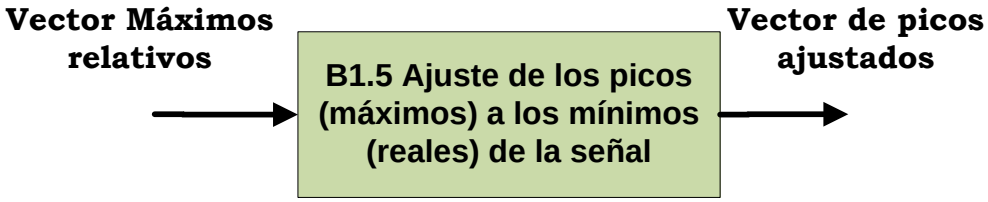

Figura 4.38: Bloque da ajuste mínimos (B1.5)

Por lo tanto, la entrada de este bloque es un vector con las magnitudes de unos máximos relativos. Como se ha comentado, una vez reajustados estos instantes de pitch a los mínimos originales, la salida del bloque, etiquetado B1.5, es un vector con la magnitud de los mínimos relativos de la señal original. Dicho vector estará rellenado con ceros en los intervalos entre las marcas de pitch (mínimos reales).

Una muestra de la salida de este bloque se muestra a continuación:

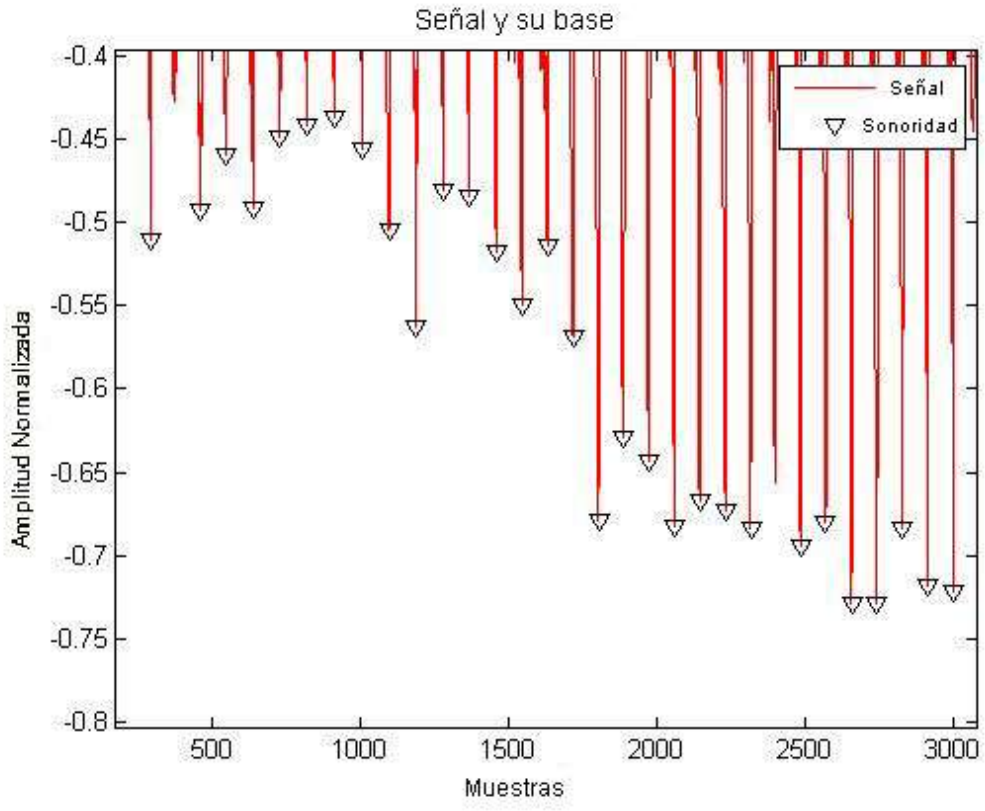

Figura 4.39: Zoom de la señal (rojo) y los mínimos de la señal

Los picos mostrados en rojo corresponden a los mínimos originales de la voz. Los intervalos intermedios estarán rellenos de ceros.

4.2.2.1.6 Eliminar mínimos por debajo del umbral (B1.6)

Una vez se consigue el vector con los mínimos relativos de la señal, se desestiman aquellos que estén por debajo de un umbral establecido. En esta primera iteración del algoritmo se establece que el umbral sea el 20% del mínimo absoluto de la señal de voz. Para la siguiente iteración dependerá de la clasificación que se realice de la señal (se detallará más adelante).

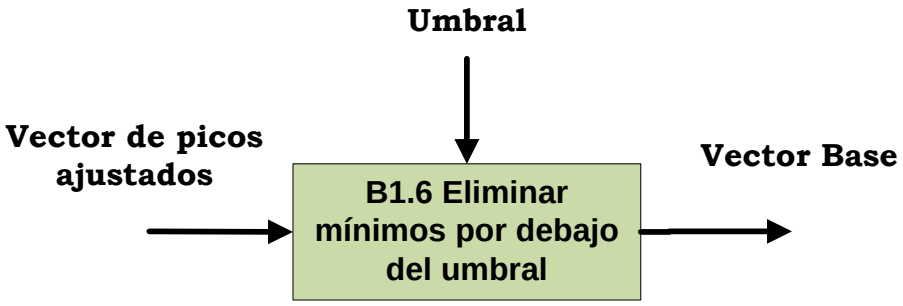

Figura 4.40: Bloque de eliminación de mínimos por debajo del umbral (B1.6)

Con lo cual, este bloque constará de dos entradas. Una de ellas, proveniente de la etapa anterior, el vector de picos ajustado a la señal original y, el otro, es la asignación del parámetro "umbral". La salida vendrá dada por un vector, que denominamos "Base", con la información de cada uno de los ciclos de voz detectados y que será la base para obtener las frecuencias instantáneas F_i y, por lo tanto, el pitch. Es evidente que con la información contenida en este vector resulta bien sencillo calcular el jitter y el shimmer aplicando las ecuaciones de la Tabla 2.3 y Tabla 2.5, respectivamente. Este es, por tanto, una de las salidas del "Algoritmo Base".

4.2.2.1.7 Obtener PrePitch (B1.7)

Como se ha comentado, una vez obtenido los instantes de pitch es inmediato calcular el pitch de la señal con las ecuaciones de la Tabla 2.1. A este pitch "intermedio" le llamaremos "Pre-pitch". Posteriormente, se realizará todo el bloque.

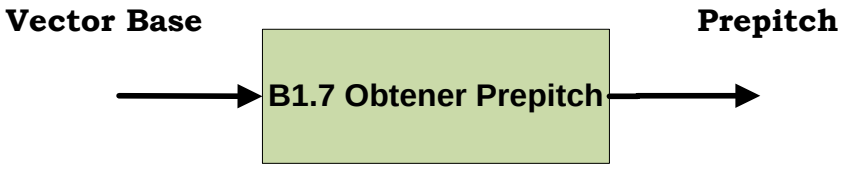

Figura 4.41: Bloque de obtención del PrePitch (B1.7)

Por lo tanto, con el vector base de entrada al bloque se obtiene la salida el PrePitch de la señal de voz. Este es el otro resultado del "Algoritmo Base".

4.2.2.2 Bloque de "Clasificación"

Se han realizado múltiples pruebas y se han comprobado los resultados para constatar que la hipótesis de partida era válida. Debido al diseño paramétrico del "Algoritmo Base" (bloques B1 o B5), éste podía ser adaptado para establecer las marcas de cualquier tipo de voz, sanas o esofágicas, en el que el problema radicaba en la *clasificación del tipo de voz* ya que se ha comprobado que la parametrización era distinta para la voz esofágica (en la que es fija) y para el resto de las voces sanas en las que es dependiente del rango frecuencial propio de la señal (ver Figura 4.31).

El bloque de clasificación, como se puede apreciar en la Figura 4.42, no es una etapa de proceso sino que es una etapa de decisión. Como se ha comentado, se pretende clasificar las voces ya que la parametrización es diferente según la señal de voz a tratar. Para ello, se toma el PrePitch como referencia y se clasifican las voces en dos bloques: sanas y esofágicas. La decisión nos conduce a dos bloques en los que se asignarán los parámetros de las voces sanas (bloque B3) y de las voces esofágicas (bloque B4) para una posterior iteración del "Algoritmo Base" (bloque B5) (Figura 4.31).

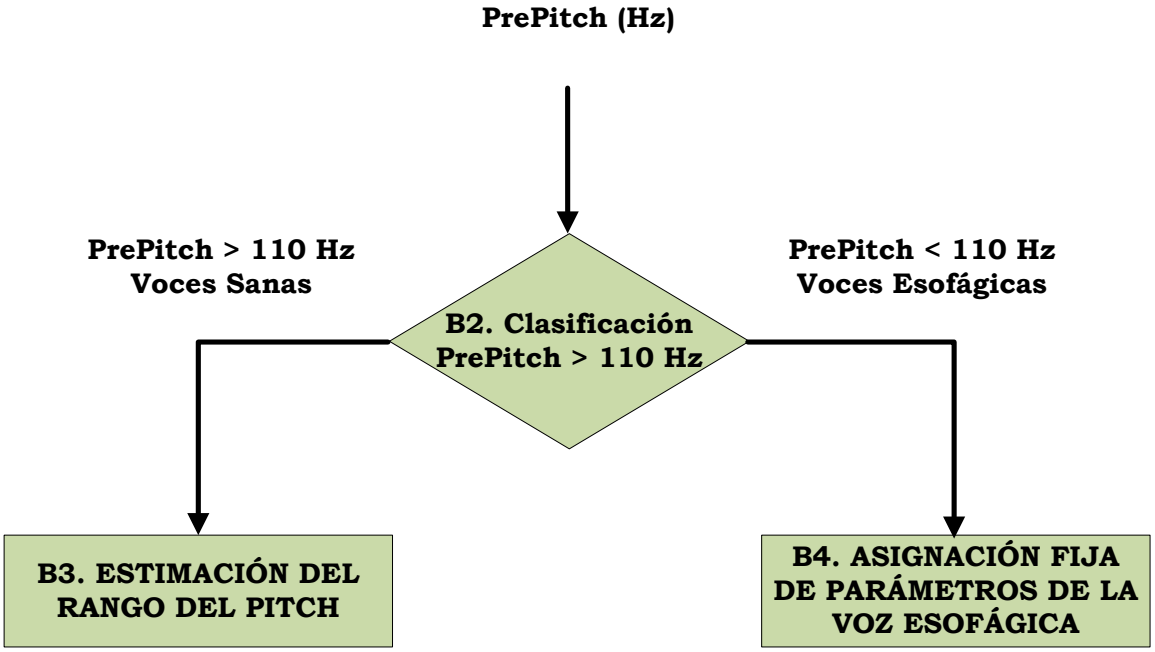

Figura 4.42: Bloque de clasificación de las voces (B2)

El límite fijado para la clasificación de las voces, en el bloque "Clasificación" está en 110 Hz (Figura 4.42) porque se ha probado que es el mejor de los clasificadores.

Por un lado, esto es porque dicho valor es cercano al límite superior real del pitch de las voces esofágicas (cuyo pitch siempre es menor que 110 Hz). Y, por otro lado, este límite funciona muy bien con voces sanas con un pitch muy bajo.

Además, dicho valor evita que haya malas medidas que puedan confundir la estimación, lo cual puede suceder ya que el rango del análisis frecuencial es tan amplio y el segundo armónico puede confundir la medida.

La etapa de clasificación podría ser desarrollada basándose en *tres criterios*: las medidas de ruido, una estimación previa del rango de pitch o quizá alguna medida experimental del "Algoritmo Base".

Las medidas de ruido acústicas están directamente relacionadas con el valor del pitch con lo cual no son válidas. Por otro lado, la estimación previa del rango del pitch se ha realizado por medio del análisis cepstral (se detallará más adelante, bloque B3) y, si bien ha mostrado ser muy válida para voces sanas, en el caso de

las voces esofágicas las componentes de ruido son tan elevadas que confunden al clasificador, lo cual invalida su uso como clasificador para las voces mencionadas. Finalmente, se ha descubierto de forma experimental que **una parametrización concreta del "Algoritmo base" puede ser empleada como clasificador** (se detallará en el bloque B4) ya que sus resultados establecen un umbral por debajo del cual las voces se consideran esofágicas.

4.2.2.3 Bloque de "Estimación del rango de pitch"

Una vez tomada la decisión de que la señal original es una voz sana, el siguiente cometido es determinar el rango de pitch de la señal para asignar los parámetros necesarios para la segunda iteración (bloque B3). Con voces sanas, la estimación del rango de pitch previo se realiza por medio del cepstrum de la señal.

El cepstrum de la señal viene dado por la ecuación:

$$C(q) = |FFT(\log \ (|FFT(señal)|))| \tag{4.15}$$

donde la escala quefrency está directamente relacionada con la escala de la frecuencia mediante la ecuación:

$$q_i = \frac{F_S}{F_i} \tag{4.16}$$

donde F_S es la frecuencia de muestreo y, F_i – q_i son la frecuencia y el quefrency relacionados en un punto determinado, respectivamente [Proakis+07].

Es bien conocido que el cepstrum ha sido utilizado para la determinación del pitch [Noll64] [Noll67]. En nuestro caso, la obtención del pitch no es muy exacta con este método y se ha utilizado para estimar el rango del pitch de la señal y, así, asignar los parámetros adecuados.

Se puede comprobar que si se transforma una señal de voz al espectro cepstral se obtiene un pico en el quefrency del pitch como se puede ver en la Figura 4.43.

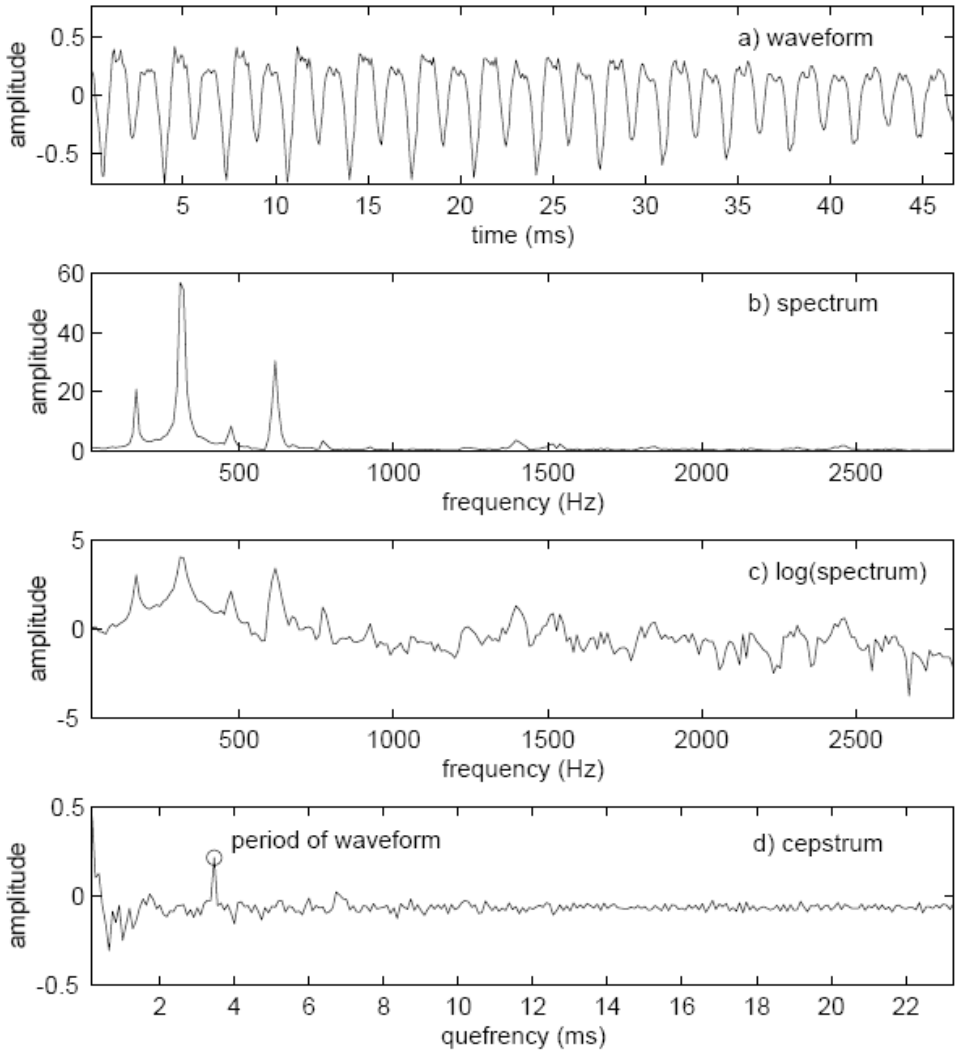

Figura 4.43: Transformación al dominio cepstral de una señal de voz

En la figura anterior se observa que la primera imagen etiquetada como a) es espectro temporal de la señal de voz. La figura b) es el espectro frecuencial de la señal, es decir, una vez realizada la transformada de Fourier (FFT). La tercera imagen, la c), muestra el logaritmo del espectro. Finalmente, se realiza de nuevo la transformada de Fourier, figura d), para obtener el cepstrum tal y como se ha descrito en la ecuación 4.15. En esta imagen viene remarcado el máximo absoluto de la magnitud en el cepstrum que indica el pitch aproximado de la señal de voz.

La entrada del bloque B3 es la señal de voz original y la salida es una asignación de parámetros dependiendo de dónde esté el rango de pitch (ver Figura 4.44). Una vez se obtiene el rango de pitch, se asignan los siguientes parámetros:

número de muestras de la ventana o trama, número de muestras de desplazamiento de la ventana, valor umbral, número de muestras mínimo para que haya un instante de pitch y número de muestras máximo para eliminar un instante de pitch. Los tres primeros parámetros se utilizan como entrada para la segunda iteración del "Algoritmo base" etiquetado como B5, los dos últimos se utilizan para el bloque "Acciones correctoras" etiquetado como B6.

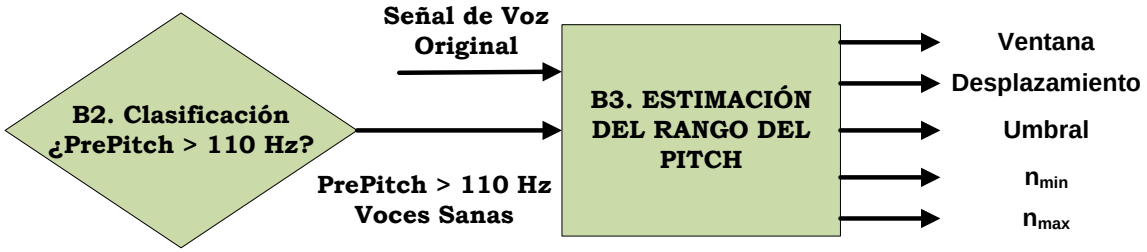

Figura 4.44: Bloque de estimación del rango de pitch (B3)

Con esta aproximación cepstral al pitch, podemos encajar todo el rango de pitch a bandas de frecuencia de 20-30Hz, en los cuales se utilizará diferentes parámetros que serán introducidos en el "Algoritmo base", en una segunda iteración, para la correcta obtención del valor del pitch.

El bloque detallado del bloque estimación del rango de pitch (B3) es el que se puede ver en la Figura 4.45.

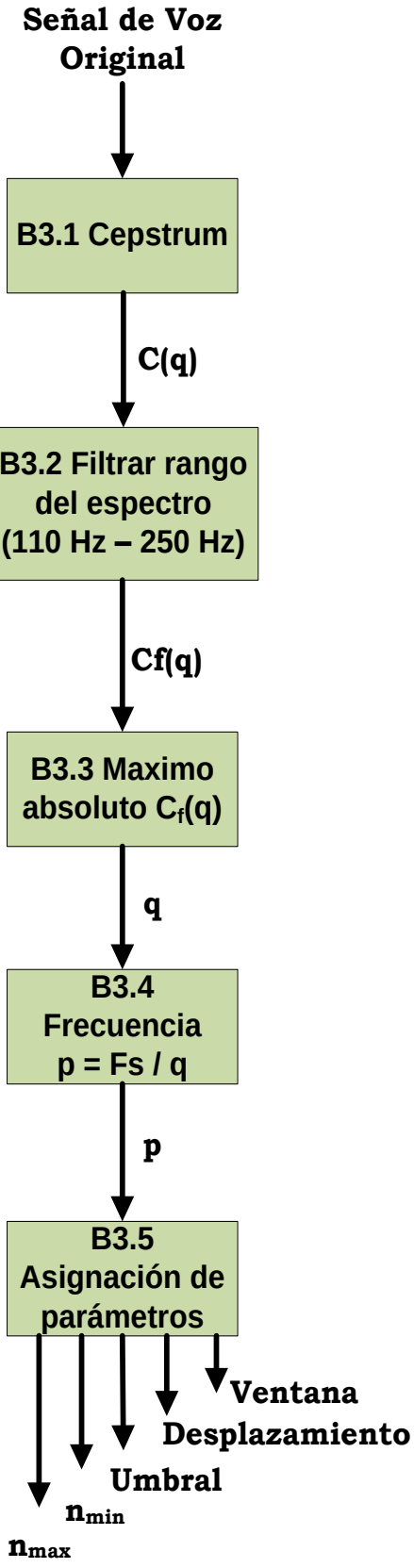

Figura 4.45: Bloque detallado de la estimación del rango de Pitch (B3)

4.2.2.3.1 Cepstrum (B3.1)

El primer paso es obtener el cepstrum de la señal voz original (ver Figura 4.46).

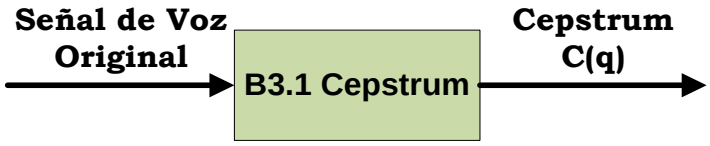

Figura 4.46: Bloque del cálculo del Cepstrum (B3.1)

Por lo tanto, la entrada de esta etapa B3.1 es la señal de voz original y la salida es el espectro del cepstrum completo. Este bloque se calcula mediante la ecuación 4.15 ya descrita.

4.2.2.3.2 Filtrar el rango del espectro del Cepstrum (B3.2)

La estimación del pitch se obtiene como el valor máximo absoluto C(q) en el rango correspondiente de 110 Hz a 250 Hz ($q_i = \frac{F_s}{250\ Hz}$ y, $q_f = \frac{F_s}{110\ Hz}$ ya que el dominio quefrency está invertido). Por lo tanto, el segundo paso es filtrar el espectro del Cepstrum en el rango de frecuencia descrito.

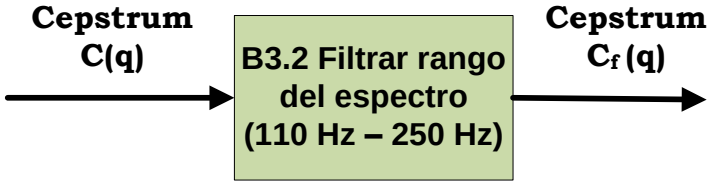

Figura 4.47: Bloque del filtro del espectro del cepstrum (B3.2)

La entrada de esta etapa es el espectro del Cepstrum completo y la salida será el espectro del cepstrum filtrado (ver Figura 4.47).

4.2.2.3.3 Obtener el máximo absoluto del espectro (B3.3)

El máximo absoluto del espectro del cepstrum filtrado es el que nos dará el valor de la medida del pitch. Este valor no es el valor de pitch exacto que estamos buscando, pero sí que se aproxima al rango de pitch donde el valor real del pitch estará. En este punto, se debe resaltar que en la categoría de voces esofágicas no

se puede realizar este análisis ya que en el rango a analizar el quefrency es muy ruidoso y la obtención del pitch sería errónea.

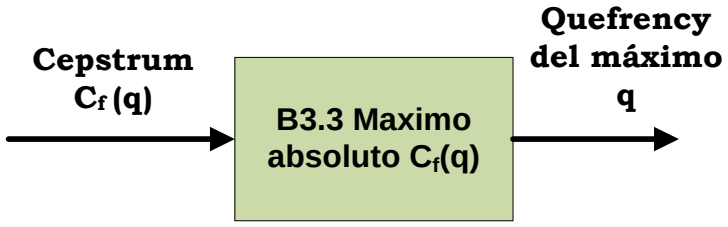

Figura 4.48: Bloque de obtención del máximo absoluto (B3.3)

La entrada de esta etapa es el espectro del cepstrum filtrado y la salida es el valor del quefrency donde esté el máximo absoluto del espectro (ver Figura 4.48).

4.2.2.3.4 Obtener la frecuencia del máximo absoluto (B3.4)

En la etapa anterior hemos obtenido el valor del quefrency y se debe relacionar con la frecuencia de ese máximo. Esta relación la da la ecuación 4.15.

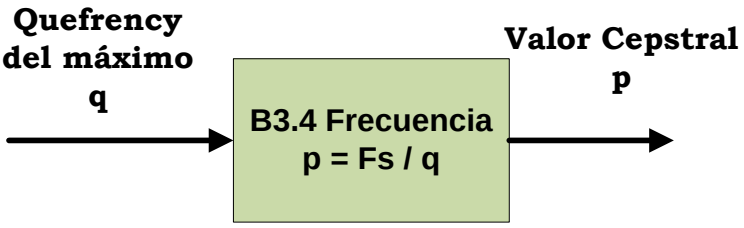

Figura 4.49: Bloque de obtención de la frecuencia del máximo absoluto (B3.4)

La entrada de esta etapa el valor del quefrency y la salida es el valor cepstral "p" (ver Figura 4.49).

4.2.2.3.5 Asignación de parámetros (B3.5)

Los parámetros que se delimitan son los que se van a utilizar para el bloque "Algoritmo Base" (B5) y "Acciones correctoras" (B6):

- **Ventana**: longitud de la ventana en muestras usado en el análisis del "Algoritmo base" (ver Figura 4.33).
- **Desplazamiento**: Número de muestras que se desplaza la ventana (ver Figura 4.33).

- **Umbral**: el valor relativo que se usa para descartar los picos con menor energía del vector de ciclo de la voz en el "Algoritmo Base".

Para la etapa de "Acciones correctores" se utilizan estos dos parámetros:

- $\mathbf{n_{min}}$ y $\mathbf{n_{max}}$: parámetros de corrección que se explicarán en el próximo apartado (se detallará más adelante).

Figura 4.50: Bloque de asignación de parámetros (B3.5)

La asignación de los parámetros de ventana, desplazamiento, umbral etc. que se introducen en el algoritmo mencionado aparecen en la Tabla 4.3. En dicha tabla "p" es el valor cepstral de la estimación del pitch y F_S es la frecuencia de muestreo de la señal. La entrada de la esta etapa es el valor cepstral y la salida son la totalidad de los parámetros descritos (ver Figura 4.50).

En la Tabla 4.3 se hace un resumen de todas las asignaciones de los parámetros que se requiere en el "algoritmo base" (B5) y en las "acciones correctoras" (B6). Hay tres tipos de asignación de parámetros: El primero, pertenece al caso en que se examina la señal por primera vez. En este caso, los parámetros asignados al "Algoritmo base" son fijos ya descritos anteriormente y se utilizan siempre los mismos.

El segundo de los casos, es aquél en el que el pre-pitch es menor del umbral tomado experimentalmente, 110 Hz. Cuando el pitch es menor del umbral seleccionado eso quiere decir que estamos analizando una señal esofágica. En este segundo caso también los parámetros asignados son fijos y se utilizan estos mismos para todas las voces esofágicas (se detallará en el próximo apartado).

Tabla 4.3: Asignación de parámetros dependiendo del rango del Pitch

Caso	**Ventana**	**Despl.**	**Umbral**	**n_{min}**	**n_{max}**
Pre-pitch 1ª iteración					
Siempre	64	10	0,2	-	-
Pre-pitch < 110 Hz					
Siempre	72	10	0,5	F_S /250	F_S /40
Pre-pitch > 110 Hz					
100 < p <110	50	10	0.2	F_S /150	F_S /40
110< p <130	50	10	0.2	F_S /(p+5)	F_S /(p-5)
130 < p <160	50	10	0.2	F_S /(p+15)	F_S /(p-16,5)
160 < p <180	50	10	0.2	F_S /(p+15)	F_S /(p-15)
180 < p <210	20	5	0.2	F_S /(p+15)	F_S /(p-25)
210 < p <230	20	5	0.2	F_S /(p+25)	F_S /(p-15)
p <230	10	5	0.2	F_S /270	F_S /220

En la tabla anterior el parámetro F_S, frecuencia de muestreo de la señal, tiene el valor de $F_S = 44{,}1\ kHz$.

El tercer y último caso, como ya se ha mencionado, es aquél en el que el pitch se aproxima en primera instancia por medio del cepstrum. En este caso estamos tratando con voces sanas y dependiendo del rango de pitch tendremos una asignación de parámetros u otra (ver Tabla 4.3). Cabe destacar que también se tiene en cuenta el caso en el que el valor cepstral "p" toma valores entre 100 Hz y 110 Hz para casos en el que el ruido puede confundir al clasificador.

4.2.2.4 Bloque de "Asignación de parámetros de las voces esofágicas"

Si en el bloque de clasificación (B2) se observa que el Pre-pitch es menor que el umbral establecido empíricamente, 110 Hz, se considera que la señal de voz original es esofágica. Como ya se ha mencionado, se ha encontrado una estimación de parámetros fija que ya ha sido expuesta en la Tabla 4.3 (el segundo de los casos).

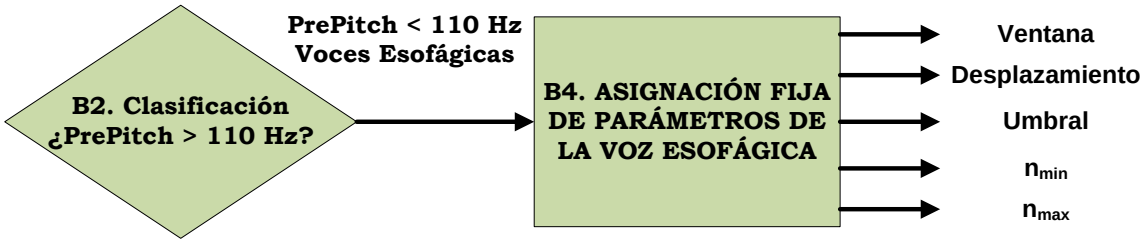

Figura 4.51: Bloque de asignación de parámetros de la voz esofágica (B4)

Esta etapa no tiene una entrada como tal sino que proviene de una decisión y la salida es el conjunto de parámetros ya mencionados en el apartado anterior.

4.2.2.5 Bloque "Algoritmo Base" (segunda iteración)

Una vez que se asignan los parámetros base, dependiendo de si la voz es sana o esofágica, tal y como se ha visto en los dos apartados inmediatamente anteriores, éstos se utilizan para la segunda iteración del "Algoritmo Base", etiquetado como B5.

Se presenta aquí el diagrama bloque "Algoritmo Base" en su segunda iteración:

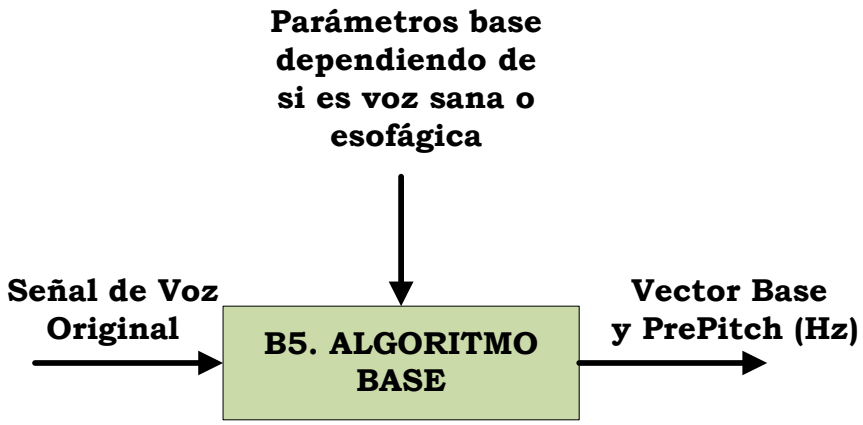

Figura 4.52: Bloque del Algoritmo base en su segunda iteración (B5)

Este bloque tiene las mismas entradas que el presentado en la Figura 4.32 ya explicado en el apartado 4.2.2.1. La diferencia es que los parámetros base ahora son diferentes dependiendo de si la voz ha sido etiquetada como sana o esofágica. Además, dichos parámetros son diferentes a los utilizados en la primera iteración (ver Tabla 4.3). La salida sigue siendo el vector denominado Base y el cálculo del Prepitch.

4.2.2.6 Bloque de "Acciones Correctoras"

Como se ha descrito hasta el momento, tres de los parámetros de que se obtienen de los bloques de "Estimación del rango de pitch" (bloque B3) para voces sanas y el de "Asignación de parámetros de las voces esofágicas" (bloque B4) para voces esofágicas se utilizan para la segunda iteración del bloque "Algoritmo base" (bloque B5). Estos parámetros son: la ventana, el desplazamiento y el umbral.

Aparte de los parámetros ya mencionados anteriormente, se introducen dos nuevos parámetros con el propósito de corregir el algoritmo descrito hasta ahora. Estos parámetros representan el mínimo (n_{min}) y el máximo (n_{max}) número de muestras permitidas entre picos consecutivos del vector base. Estos parámetros ya se han presentado en la Tabla 4.3 pero aún no han sido descritos en profundidad.

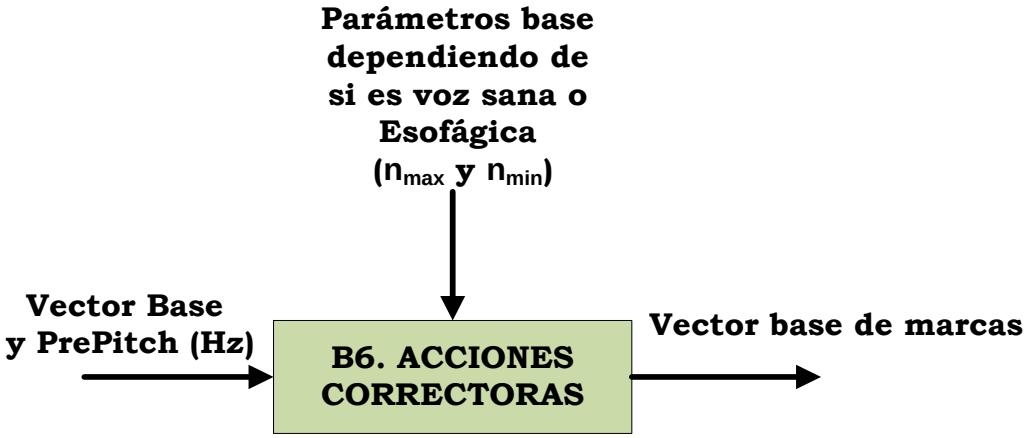

Figura 4.53: Bloque de Acciones correctoras (B6)

La entrada del bloque de "Acciones correctoras" tiene como entrada, por un lado, la salida del bloque B5 (segunda iteración del "Algoritmo base"), el vector base y el prepitch y, por otro lado, el número máximo y mínimo de muestras (n_{min}, n_{max}) provenientes de los bloques B3 o B4 ya mencionados (ver Figura 4.53).

En términos generales, el rango de frecuencias de la voz humana, en condiciones normales, oscila entre 40 Hz y 250 Hz incluyendo aquí las voces patológicas más graves como son la voz esofágica. Así que, introduciendo esta peculiaridad de la voz, se puede entender de forma intuitiva que el algoritmo va a mejorar.

La propia variabilidad de las medidas de voz por las cuales se observaban ciertos errores sistemáticos y el afán de maximizar la precisión del algoritmo han conducido a desarrollar ciertos **mecanismos de corrección** como, por ejemplo: eliminación de sub-harmónicos, detección de silencios y marcas inadvertidas e interpolación de marcas.

El primer mecanismo *reduce los errores de baja frecuencia* evitando considerar regiones de silencio o donde puede haber pasado una marca inadvertida. El segundo, *reduce los errores de alta frecuencia* debidos a marcas muy cercanas a las reales. Los dos sub-bloques de corrección se pueden observar en la Figura 4.54.

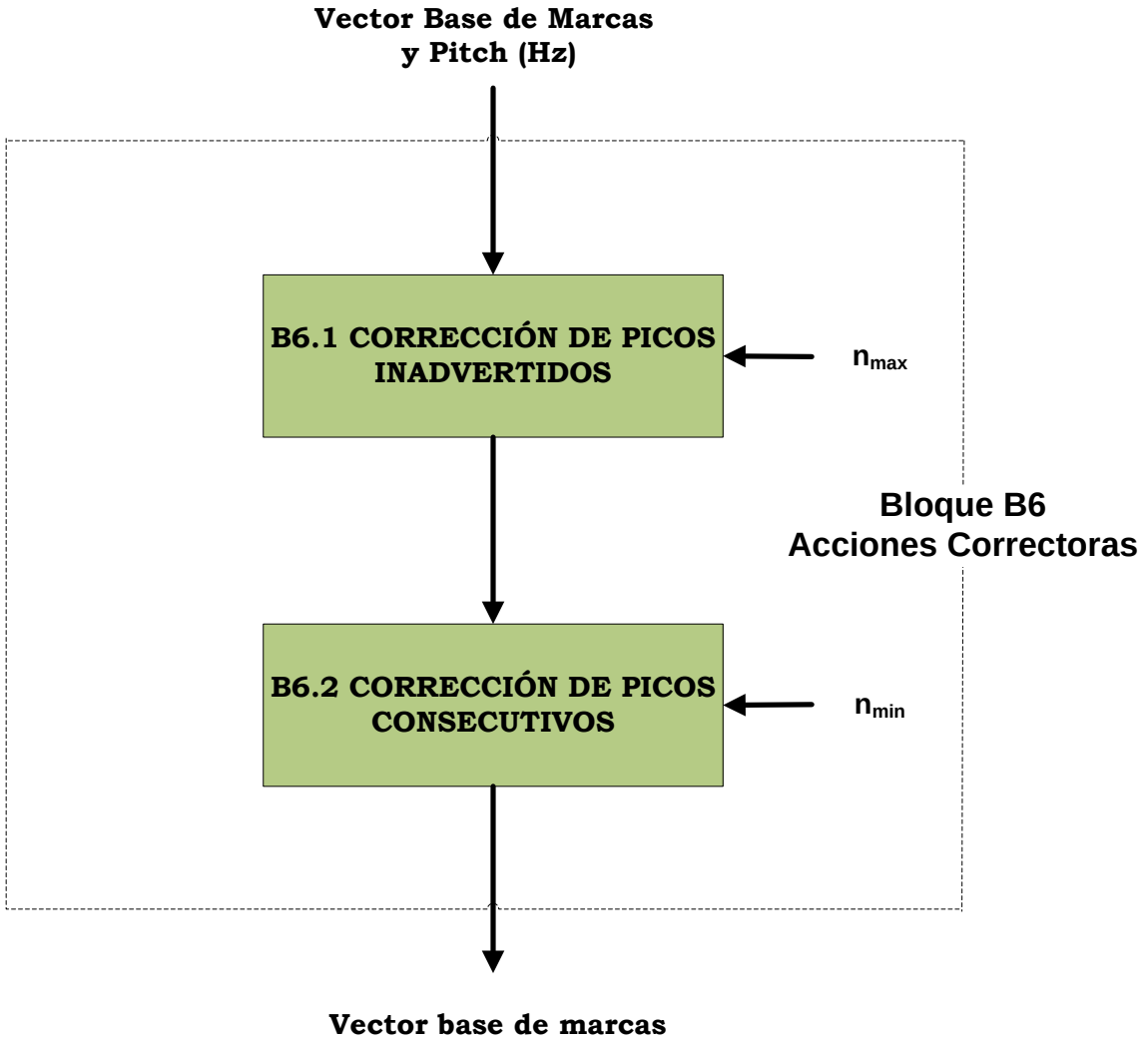

Figura 4.54: Diagrama de bloques de las acciones correctoras detallado

4.2.2.6.1 Corrección de picos inadvertidos (B6.1)

El algoritmo descrito hasta ahora tiene dos inconvenientes o puntos débiles. El primero está relacionado con el parámetro "umbral". Si el umbral elegido no tiene el nivel adecuado puede que haya picos que no se detecten y, por lo tanto, esto afecta al pitch. Se puede decir en este caso el algoritmo ha perdido el seguimiento del pitch de la voz por un momento.

Esto produce un incremento de la distancia en muestras entre dos picos consecutivos, lo cual conduce a que el algoritmo obtenga pitch instantáneos de valores muy bajos. Ahora bien, esto no quiere decir que la voz tenga

componentes de pitch bajos sino que deben ser corregidos. Para ello, se introduce el parámetro n_{max} en el bloque "Corrección de picos inadvertidos", etiquetado como B6.1 (ver de la Figura 4.54). Por lo tanto, si la distancia en muestras entre dos picos consecutivos excede de un valor dado n_{max} dicho pitch no será tomado en cuenta.

4.2.2.6.2 Corrección de picos consecutivos (B6.2)

Por otro lado, otro de los puntos débiles del algoritmo es la parametrización del enventanado para obtener el pitch entre dos picos consecutivos. Este proceso puede obtener malos resultados, normalmente debido a valores pequeños del parámetro "ventana".

Este hecho produce que el algoritmo tome dos valores de picos como instantes de pitch en una misma región. Esto sucede cuando el valor del máximo absoluto de la ventana coincide con el final de la misma y con el principio de la siguiente ventana consecutiva. Esta situación produce el efecto contrario al caso anterior. Cuando sucede este hecho, la distancia en muestras de dos picos consecutivos es muy pequeña y, consecuentemente, produce instantes de pitch elevados. Por lo tanto, uno de estos picos debe ser eliminado del vector "base" que da como resultado el "Algortimo base" de la Figura 4.52. En consecuencia, si el número de muestras entre dos picos consecutivos es menor que el parámetro $n_{min,}$ (ver "Corrección de picos consecutivos" de la Figura 4.54) uno de los picos deberá ser extraído del vector "base" donde aparecen la magnitud en la que se encuentra el máximo relativo de la señal de voz.

Aparte de corregir estos dos tipos de errores del algoritmo principal, este bloque de acciones correctivas intenta aumentar la precisión del algoritmo ajustando los parámetros de la banda donde se supone que va a estar el pitch.

Esta corrección es un tipo de filtrado de componentes de pitch de tal manera que si se espera tener un rango de pitch de 200 Hz, obtenido con el análisis cepstral descrito en la Figura 4.45 y parametrizado en la Tabla 4.3, entonces se puede

ajustar los parámetros del algoritmo principal para que "encuentre" instantes pitch cercanos a ese rango, mejorando la precisión del algoritmo. De esta manera, se aumenta la precisión de los parámetros asignados en la Tabla 4.3. Una vez realizado el proceso la tabla de asignación de parámetros ha quedado ajustada.

Por otro lado, aquellas voces con un pre-pitch menor de 110 Hz, es decir, las voces esofágicas, la estimación de los parámetros a utilizar en el "Algoritmo base" es prácticamente directa. Los parámetros de la Tabla 4.3 para las voces esofágicas y con un pitch muy bajo se utilizan directamente en el "Algoritmo base" y el pitch se extrae directamente. En esta categoría solamente se realiza una corrección leve de cara a comprobar que los componentes instantáneos de pitch están dentro del rango de la voz humana (40 Hz - 250 Hz).

4.2.2.7 Bloque de "Cálculo de parámetros

En este bloque se calculan los parámetros del pitch, jitter, shimmer y HNR a partir del vector base obtenido. Las entradas de esta etapa son el vector base de marcas y la señal de voz original de cara a obtener el valor de las amplitudes en dichas marcas. Con las marcas de de voz se obtiene fácilmente el pitch (Tabla 2.1), el jitter (Tabla 2.3), shimmer (Tabla 2.5) y HNR (apartado 2.3.4) tal y como se ha descrito en el apartado 2. Las salidas de este bloque son los valores de pitch, jitter, shimmer y HNR.

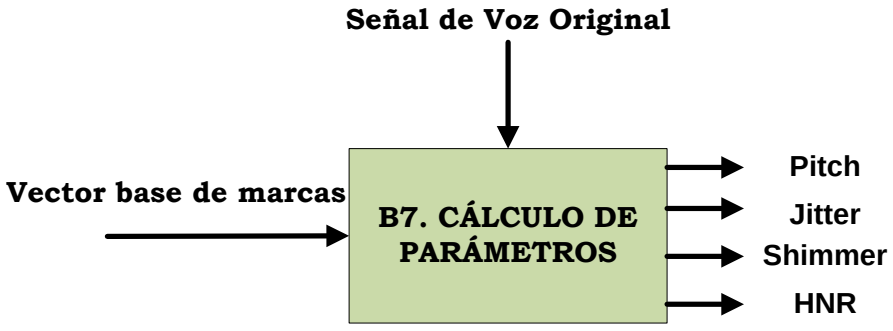

Figura 4.55: Bloque de "Cálculo de parámetros"

Cabe destacar, que los valores que se obtienen son los ya corregidos.

RESULTADOS

5. RESULTADOS

"Insanity: doing the same thing over and over again and expecting different results"

Albert Einstein

En este capítulo se procederá a mostrar los resultados obtenidos, tanto numéricos como gráficos, del algoritmo presentado en el capítulo cuatro de esta tesis: "Diseño del Algoritmo". Tal y como se expone en dicho apartado, se presentarán los resultados relativos a la etapa de "Mejora de la Calidad de la Voz Esofágica" (bloque A) y "Mejora de la Parametrización de la Voz Esofágica" (bloque B).

En la primera etapa o bloque (A), es decir, en la de mejora de la voz esofágica, se hará especial hincapié en los parámetros del Shimmer y la relación armónico ruido (HNR).

En la segunda etapa o bloque (B), la de parametrización de la voz esofágica, de cara a corroborar los resultados se ha utilizado una base de datos de voces tanto sanas como esofágicas. Para ratificar la validez del algoritmo y así certificar la hipótesis de partida se han utilizado cuatro parámetros de la voz, unos de los más relevantes: pitch, jitter, shimmer y HNR.

Por todo ello, se describirán los resultados en dos apartados independientes para evaluar de forma más clara los efectos de cada uno de los algoritmos presentados en el capítulo anterior.

En primer lugar, en este capítulo, se abordan los aspectos más destacables sobre la implementación de los algoritmos y, en segundo lugar, la fase de evaluación, donde se describen las pruebas realizadas para la validación de la hipótesis.

5.1 CONSIDERACIONES PREVIAS

En este punto se describe el entorno software y el soporte hardware utilizado para el desarrollo de esta tesis, así como el software desarrollado paralelamente a la realización de este trabajo.

5.1.1 Entorno de desarrollo

El lenguaje de desarrollo utilizado para implementar los algoritmos diseñados ha sido Matlab 7.6.0.324 (R2008a).

Las voces utilizadas son las que se han descrito en el capítulo tres de esta tesis. Todos el procesado de estas señales de voz se ha realizado utilizando Matlab. No obstante, algunos de estos algoritmos se han implementado en el programa considerado como la versión libre de Matlab, GNU Octave 2.1.50, de cara a poder implementar un entorno de desarrollo que mida de forma automática cualquier tipo de voz. Este es el software que se ha utilizado para medir las voces del algoritmo de mejora de la voz esofágica.

No obstante, de cara a validar la parametrización realizada para medir las voces, se ha realizado pruebas adicionales con otras voces: sanas y esofágicas, y se han contrastado los resultados con el Multi-Dimensional Voice Program (versión MDVP 2.7.0).

5.1.2 Hardware utilizado

En esta investigación no ha sido necesario hardware específico, excepto el micrófono utilizado para realizar las capturas de las voces esofágicas en la Asociación Vizacaína de Laringectomizados.

El desarrollo de esta tesis se ha centrado en el procesado de las voces proporcionadas por el especialista una vez grabadas. Al no realizarse en tiempo real, no es necesario hardware de alto rendimiento de características especiales.

Los requisitos mínimos necesarios para el desarrollo y puesta en marcha de los algoritmos desarrollados son los siguientes:

Tabla 5.1: Especificaciones técnicas del hardware utilizado

Ítem	Característica
Ordenador	Acer Aspire 4820TG
Procesador	Intel® Core™ i5 CPU M480 @2,67GHz
Memoria RAM	4,00 GB
Sistema Operativo	Windows 7 Professional

Como se puede observar no es necesario ningún requisito específico de hardware avanzado para poder procesar las voces.

5.2 EVALUACIÓN DE LOS RESULTADOS

En esta sección se describen los resultados obtenidos en las etapas de "Mejora de la Calidad de la Voz Esofágica" (bloque A, sección 5.2.1) y "Mejora de la Parametrización de la Voz Esofágica" (bloque B, sección 5.2.2).

Dentro del bloque A, se mostrarán los resultados parciales de las etapas: Algoritmo de la Transformada Wavelet (A1, sección 5.2.1.1), Filtrado Kalman (A2, sección 5.2.1.2) y Estabilización de Polos (A3, sección 5.2.1.3).

Dentro del bloque B, se realizarán pruebas con voces sanas y esofágicas para dar validez algoritmo automático a todo tipo de voces. Como ya se ha comentado se compararán los resultados con un paquete de software comercial, el MDVP.

5.2.1 Pruebas del Algoritmo de "Mejora de la Calidad de la Voz Esofágica"

Como ya se ha mencionado, en este primer bloque A de pruebas se van presentar tres sub-bloques de resultados. A continuación se detallarán todos los resultados parciales de cada etapa, incluyendo imágenes de señales de voz para mayor claridad del proceso realizado.

5.2.1.1 Pruebas de la Etapa de la Transformada Wavelet

Con la base de datos de treinta voces esofágicas que se dispone se ha implementado el algoritmo de la transformada wavelet. La entrada este algoritmo es una señal de voz esofágica a una frecuencia de muestreo en la entrada y en la salida de F_{SIN}=44,1 kHz. La wavelet madre utilizada para esta prueba es la "bior 6.8" y el nivel utilizado para descomponer la señal original es 7, tal y como se ha expuesto en el capítulo 4.

En esta etapa podemos mostrar qué señales obtenemos en cada una de las sub-etapas ya que se aplican las transformaciones al conjunto de la señal. Se toma una de las voces esofágicas que se muestra a continuación:

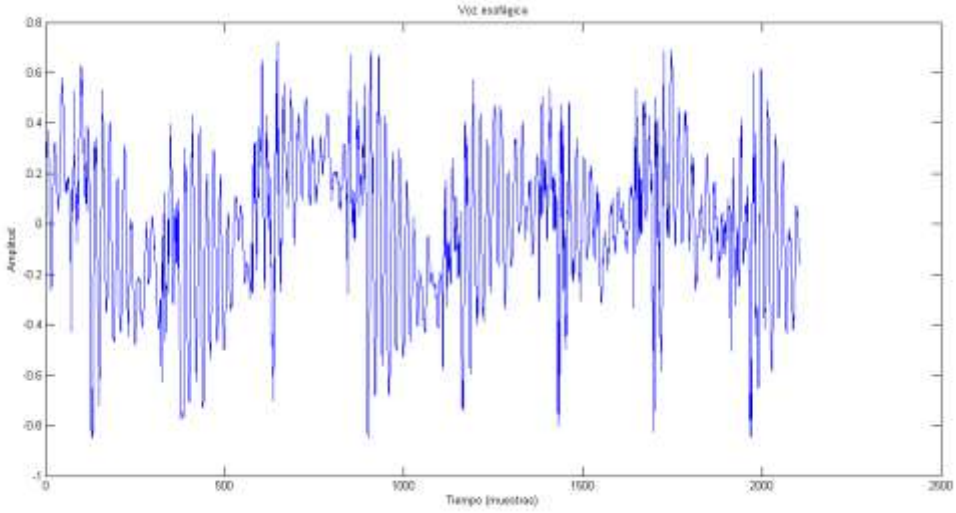

Figura 5.1: Señal de voz esofágica original "A1.wav".

Lo primero que se realiza en esta etapa es un re-muestreo de la señal a 12,8 kHz (etapa A1.1). Esta señal re-muestreada se muestra a continuación:

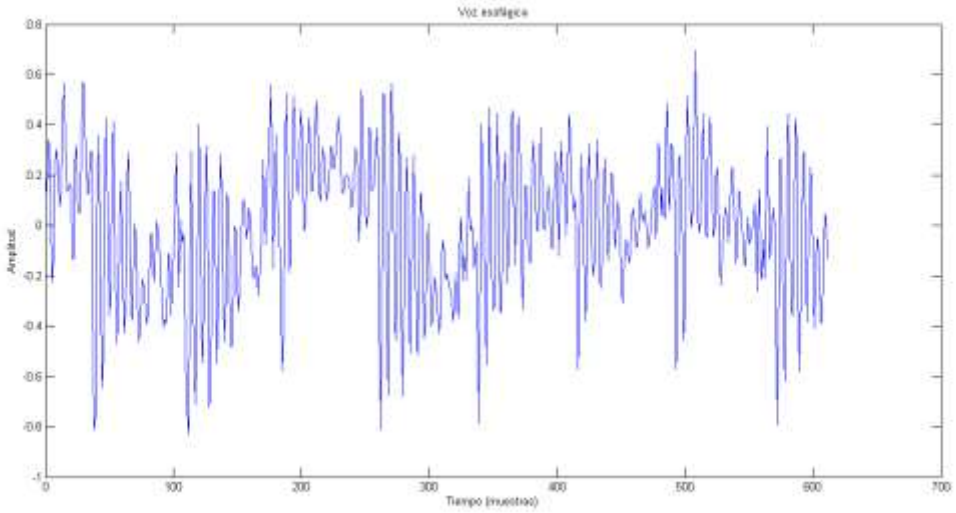

Figura 5.2: Señal de voz esofágica re-muestreada

Tal y como se puede apreciar no existen demasiadas diferencias entre una y otra, obviamente.

Después, se aplica la Trasformada Wavelet Discreta (DWT) y, de este bloque se obtienen ocho señales de coeficientes que se muestran en sendas figuras (etapa A1.2):

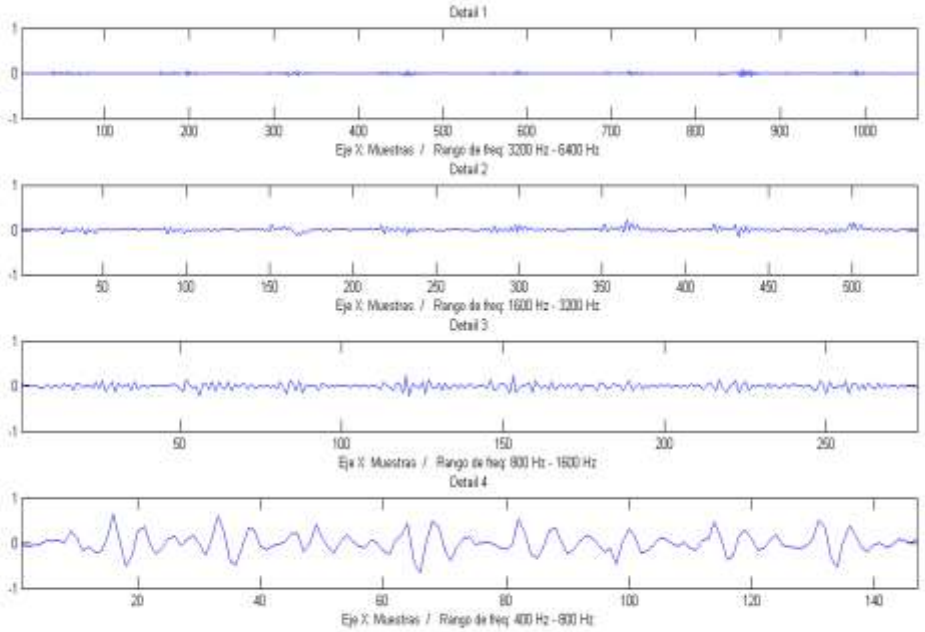

Figura 5.3: Detalles de 1 a 4 de la DWT

Como se aprecia en la Figura 5.3 el eje de abscisas están representados los coeficientes wavelet. Los otros detalles se presentan a continuación:

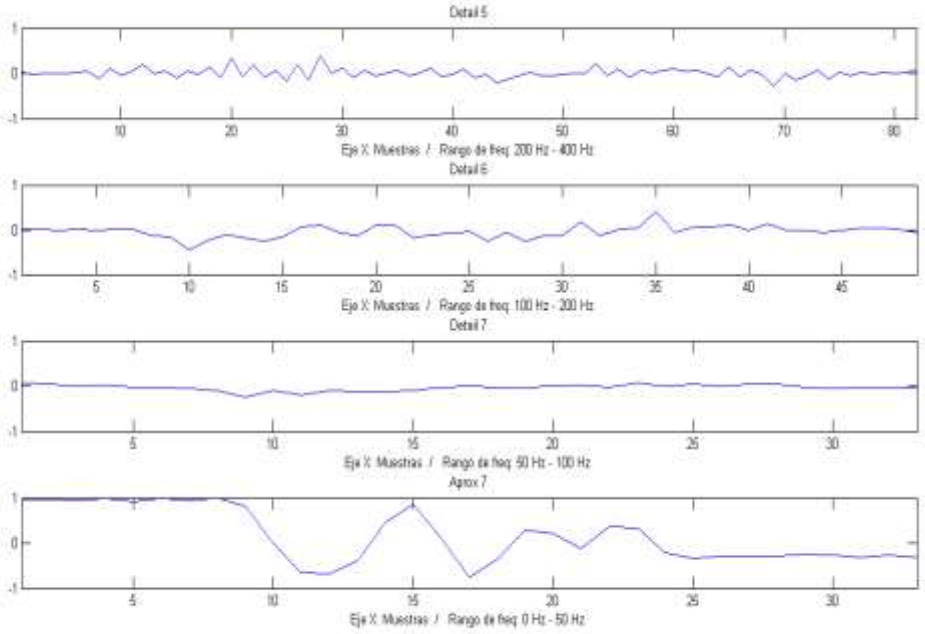

Figura 5.4: Detalle de 5 a 7 y la aproximación 7 de la DWT

Se puede apreciar en las distintas gráficas que el número de muestras se va dividiendo por dos de un detalle a otro. Esto se puede apreciar en el rango del eje de las abscisas. Posteriormente, se obtiene estas mismas señales en el dominio temporal con los mismos rangos de frecuencias. Estas señales se pueden observar en sendas figuras a continuación:

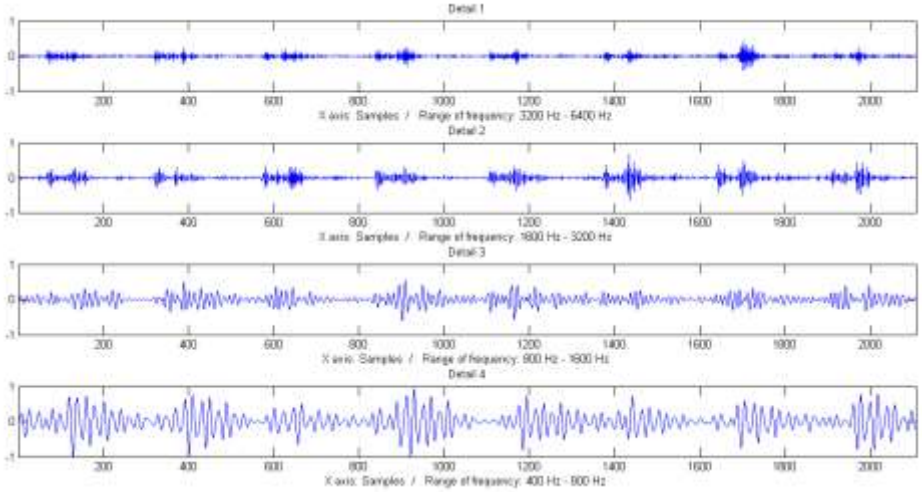

Figura 5.5: Detalles de 1 a 4 de la DWT en el dominio temporal

Los detalles restantes y la aproximación se muestran a continuación:

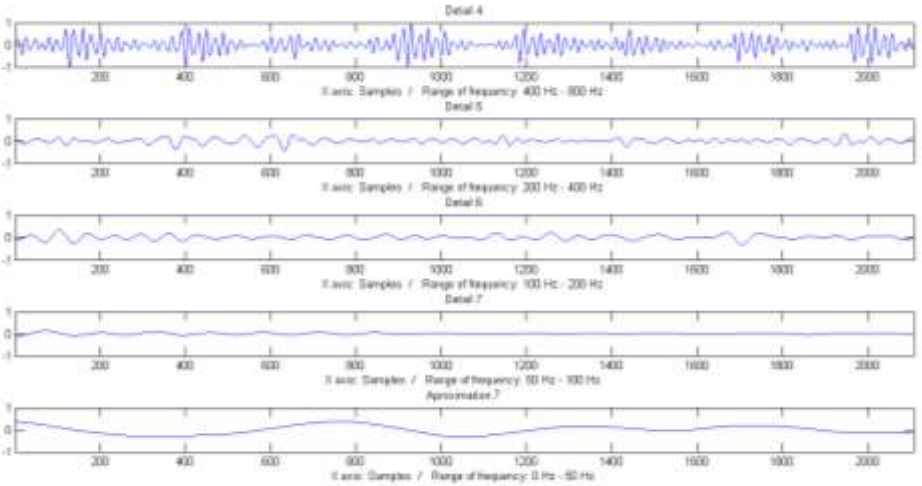

Figura 5.6: Detalles de 5 a 7 y aprox. de la DWT en el dominio temporal

Una vez que se obtienen las señales en el dominio temporal, se elimina el ruido de baja frecuencia como se puede observar en la siguiente figura (etapa A1.3):

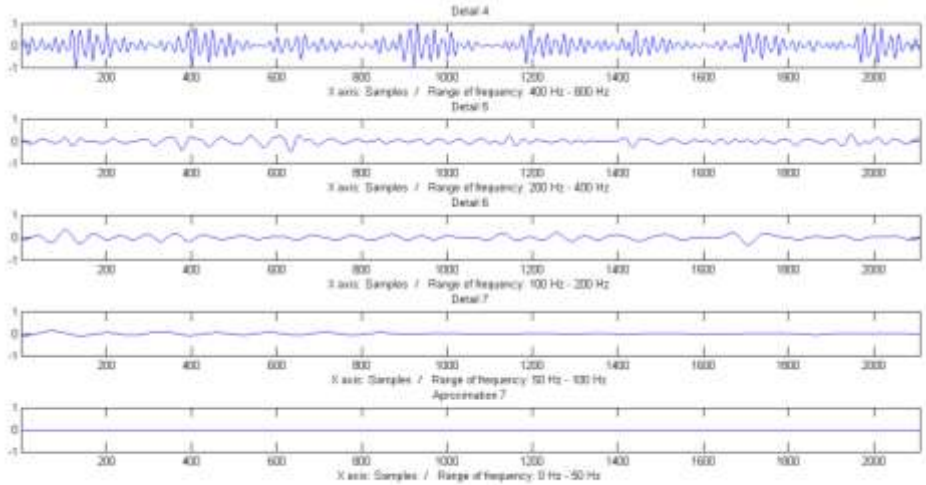

Figura 5.7: Eliminación del ruido de baja frecuencia

Para obtener la señal reconstruida es necesario realizar la suma de todos los detalles d1(n),…,d7(n) y la aproximación a7(n) (etapa A1.4). Esta señal reconstruida una vez procesada se puede observar en la siguiente figura:

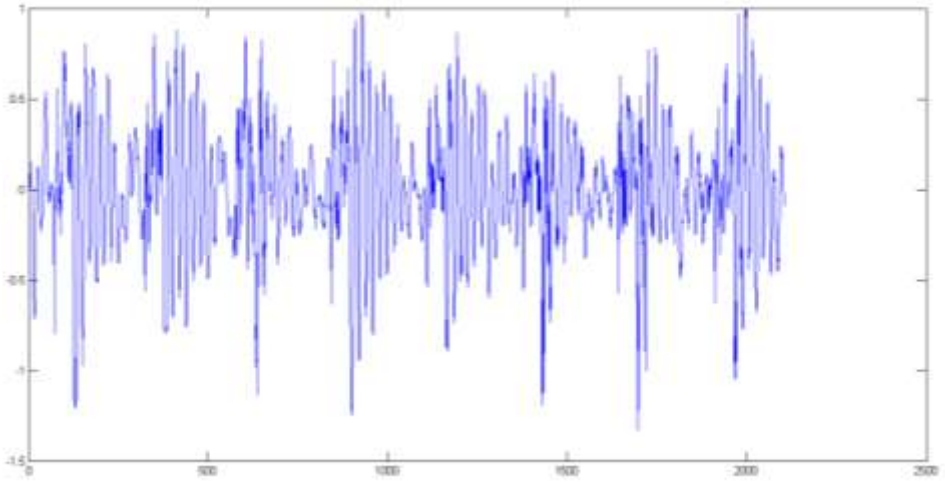

Figura 5.8: Voz esofágica procesada de la etapa A1

Se puede apreciar claramente que ya no existe el ruido de baja frecuencia. Se puede comparar esta voz esofágica procesada con la de la Figura 5.1 o con la de la Figura 5.2. Esto hace que la amplitud de los picos en los instantes de pitch sea más homogénea y, por lo tanto, disminuya el shimmer de la señal. Finalmente, para que la señal de salida tenga la misma frecuencia de muestreo que la original, se realiza el re-muestreo inverso (etapa A1.5).

Tabla 5.2: Resultados de la etapa de Transformada Wavelet

	Voces originales		**Procesadas (DWT)**	
	Shimmer (dB)	**HNR (dB)**	**Shimmer (dB)**	**HNR (dB)**
A1	0,594	-2,098	0,106	-0,531
A2	0,468	-6,959	0,227	-1,762
A3	2,734	-7,070	0,254	-6,053
A4	0,573	-7,979	0,372	-5,713
A5	1,323	-6,641	0,959	-5,756
A6	0,796	-0,802	0,521	1,153
A7	0,409	-9,191	0,175	-9,073
A8	0,342	-8,481	0,186	-8,223
A9	0,673	-2,526	0,556	-2,907
A10	0,339	-7,851	0,171	-5,359
A11	1,500	-7,298	0,816	-6,279
A12	0,412	-5,425	0,289	-5,054
A13	0,909	-5,700	0,519	-5,057
A14	0,868	-2,292	0,346	-0,651
A15	0,416	-6,480	0,123	-5,248
A16	0,230	-5,687	0,264	-3,651
A17	0,359	-8,557	0,289	-6,509
A18	0,710	-7,735	0,413	-4,544
A19	2,010	-7,370	1,97	-6,983
A20	0,343	-6,570	0,124	-6,388
A21	0,863	-5,070	0,477	-5,002
A22	0,997	-5,040	0,272	-4,632
A23	1,980	-3,644	0,712	-3,257
A24	0,494	-5,010	0,275	-4,474
A25	1,164	-6,419	0,541	-6,861
A26	2,069	-9,309	0,599	-4,576
A27	2,002	-9,191	2,073	-8,658
A28	2,133	-6,638	1,151	-6,962
A29	1,461	-3,772	0,208	-3,981
A30	2,772	-7,796	1,566	-6,898

En la Tabla 5.2 se muestran las medidas realizadas de los parámetros Shimmer y HNR para las voces de entrada y salida de la etapa de la transformada wavelet.

Como se puede apreciar en la Tabla 5.2 el Shimmer ha disminuido en 0,513 dB de media. Esto supone una mejora considerable en este parámetro y en la calidad de la voz en general. De dicha tabla se puede extraer que las voces que sufren una mayor mejora en el Shimmer son las etiquetadas como A3, A23, A26, A29 y A30. Esto se aprecia con mayor claridad en la siguiente figura:

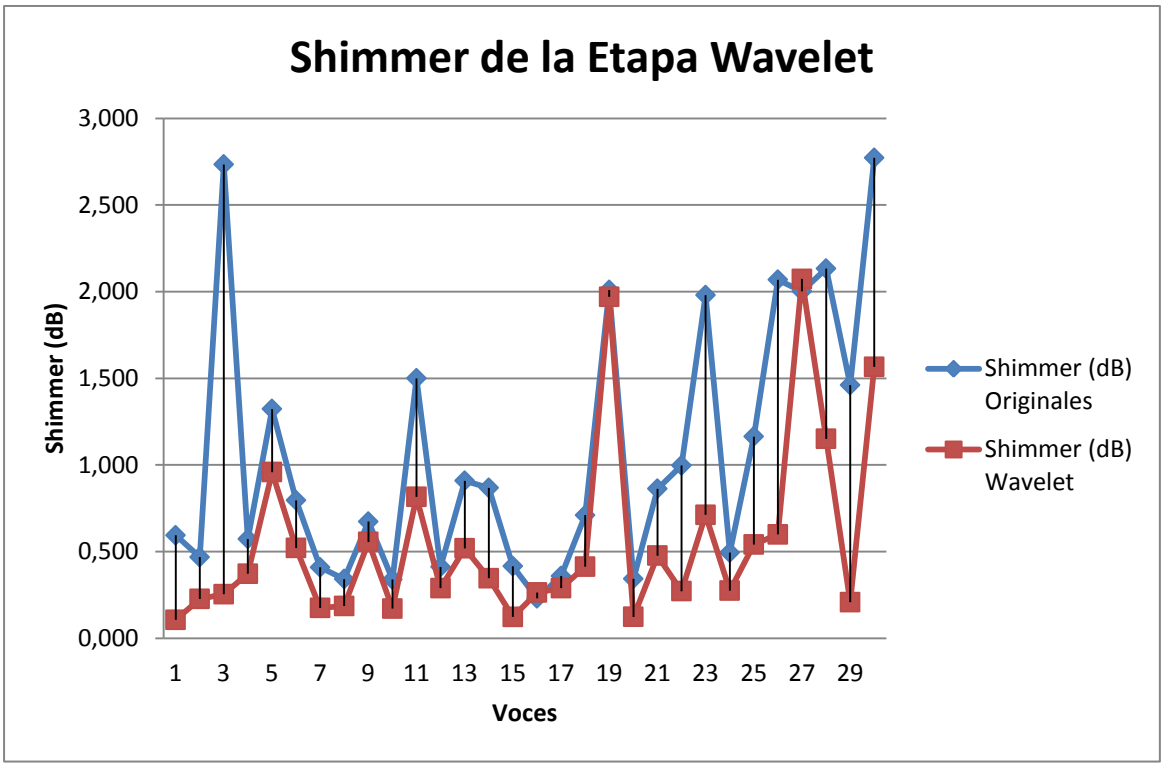

Figura 5.9: Comparación del Shimmer en la primera etapa.

Como se puede apreciar claramente, el Shimmer de las voces después de ser procesadas por la primera etapa utilizando la técnica de la transformada wavelet (en rojo) es menor que el Shimmer de las voces originales (en azul), con lo que esto supone una mejora en la calidad de la voz.

Si analizamos los datos desde un punto de vista estadístico, se debe comentar que los datos del shimmer no cumplen el criterio de normalidad como ya se ha comentado en el apartado 4, es decir, no podemos asumir normalidad en los datos. Esto determina en cierta manera el estudio estadístico a realizar. Por lo

tanto, el estudio "T-student" [Park+11] no es válido para un distribución de datos no normalizada. De cara a conocer la normalidad de los datos se han realizado dos pruebas: test de Kolmogorov-Smirnov [Fasano+87] y el de Shapiro-Wilk [Shapiro65]. Ambas pruebas han dado como resultado que la distribución de los datos del shimmer, que se tienen en cuenta en esta etapa, no están normalizadas. La significancia de los datos es menor que 0,001 lo cual indica que no se pueden asumir la normalidad de los datos. Una vez que no podemos asumir la normalidad de los datos, se realiza la prueba Wilcoxon [Wilcoxon45] para comparar los datos. Esta prueba nos muestra que los datos del shimmer original y los datos procesados tras la primera etapa no son iguales. La significancia de la prueba es menor que 0,001 ($p<0,0001$) y, por tanto, se rechaza la hipótesis nula que dice: "La mediana entre los datos originales y los datos obtenidos tras la etapa wavelet son iguales". Este resultado con la Figura 5.9 evidencia que existe una mejora en el parámetro shimmer.

En la siguiente figura se muestra el comportamiento del HNR en esta etapa:

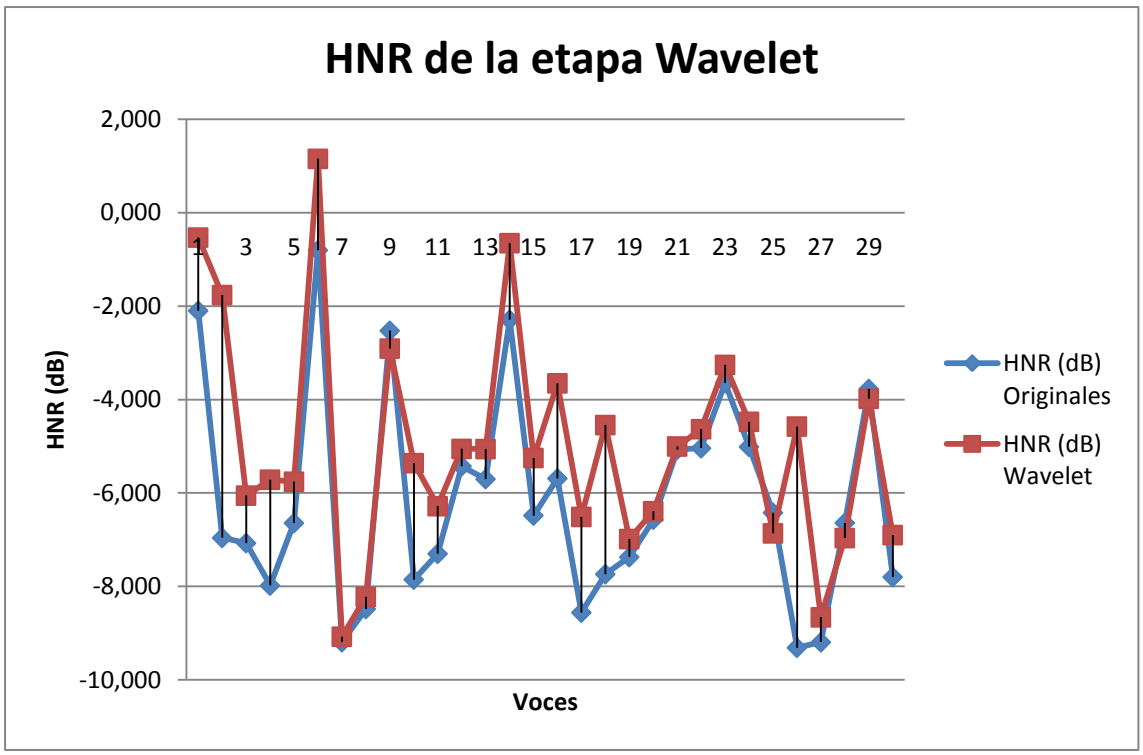

Figura 5.10: Comparación del HNR en la primera etapa.

El parámetro HNR también ha experimentado una mejora de 1,157 dB debido a la eliminación del ruido de baja frecuencia. De la Tabla 5.2 se puede observar que las voces etiquetadas como A4, A6, A16 y A18 son las voces que experimentan una mayor subida en el HNR. Estas voces son las etiquetadas como A9, A25, A28 y A29.

Como se puede apreciar claramente, el HNR de las voces después de ser procesadas por la primera etapa utilizando la técnica de la transformada wavelet (en rojo) es mayor que el HNR de las voces originales (en azul).

Desde el punto de vista estadístico, se debe comentar que los datos del HNR de las voces esofágicas originales cumplen el criterio de normalidad, es decir, podemos asumir la normalidad de los datos. De cara a conocer la normalidad de los datos se han realizado dos pruebas: test de Kolmogorov-Smirnov [Fasano+87] y el de Shapiro-Wilk [Shapiro65]. Ambas pruebas han dado como resultado que la distribución de los datos del HNR para la voz esofágica sí está normalizada y la significancia de los datos es mayor que el 5%, lo cual indica que se pueden asumir la normalidad de los datos en esta etapa. Por lo tanto, podemos utilizar la prueba estadística del "T-student" [Park+11].

Una vez que podemos asumir la normalidad de los datos, se realiza la prueba "T-student" [Park+11] para comparar los datos. Esta prueba nos muestra que los datos que no son iguales: los originales y los datos tras el procesado de la primera etapa. La significancia es menor a un 5%, concretamente, $p<0,0001$. Esto nos indica que se rechaza la hipótesis nula, en la que se dice que las medias de los datos, originales y procesadas, son iguales.

Esto junto con la Figura 5.10 presentada anteriormente nos muestra que en esta etapa se produce una mejora del parámetro HNR.

5.2.1.2 Pruebas de la Etapa del Filtrado de Kalman

En esta etapa, se recogen las voces procesadas en la etapa wavelet y se procesan en esta segunda etapa mediante el filtrado de Kalman. En esta etapa se procesa la señal muestra a muestra con lo que no tiene sentido realizar una exposición gráfica de cada una de las sub-etapas tal y como se ha realizado en el apartado anterior. Por tanto, se realizará una exposición del resultado global de la etapa.

En la Tabla 5.3 se muestra los valores de HNR medidos en decibelios antes y después de aplicar el filtro de Kalman utilizando los instantes de silencio de la voz esofágica como ruido de medida. Esta es la principal novedad de este algoritmo, escoger el ruido de medida como el ruido en los instantes de silencio de la voz esofágica.

En esta tabla se compara los valores obtenidos en la etapa anterior (etapa de procesado wavelet) con los obtenidos en esta etapa tanto para el Shimmer como para el HNR. En esta etapa se hace especial hincapié en el parámetro HNR.

Como se puede observa en la tabla, los valores que mejores resultados presentan con respecto al HNR son los etiquetados como A6, A7, A11-A15 y A27-A29 ya que éstos son los que mayor incremento presentan en decibelios. La mejora media presentada en esta etapa en el HNR es de 1,449 dB.

Tabla 5.3: Resultados de la etapa Filtrado de Kalman

	Procesadas (DWT)		Procesadas (Kalman)	
	Shimmer (dB)	**HNR (dB)**	**Shimmer (dB)**	**HNR (dB)**
A1	0,106	-0,531	0,123	1,501
A2	0,227	-1,762	0,215	0,686
A3	0,254	-6,053	0,481	-5,987
A4	0,372	-5,713	0,316	-5,202
A5	0,959	-5,756	0,854	-5,621
A6	0,521	1,153	0,554	3,484
A7	0,175	-9,073	0,209	-6,354
A8	0,186	-8,223	0,213	-6,539
A9	0,556	-2,907	0,504	-2,111
A10	0,171	-5,359	0,185	-4,650
A11	0,816	-6,279	0,796	-5,106
A12	0,289	-5,054	0,296	-3,908
A13	0,519	-5,057	0,562	-4,200
A14	0,346	-0,651	0,327	0,812
A15	0,123	-5,248	0,149	-2,846
A16	0,264	-3,651	0,222	-2,629
A17	0,289	-6,509	0,278	-4,154
A18	0,413	-4,544	0,461	-2,166
A19	1,97	-6,983	1,972	-6,013
A20	0,124	-6,388	0,139	-6,156
A21	0,477	-5,002	0,425	-3,035
A22	0,272	-4,632	0,314	-3,623
A23	0,712	-3,257	0,768	-2,442
A24	0,275	-4,474	0,251	-4,284
A25	0,541	-6,861	0,604	-5,723
A26	0,599	-4,576	0,572	-2,839
A27	2,073	-8,658	1,923	-6,488
A28	1,151	-6,962	1,356	-3,987
A29	0,208	-3,981	0,442	-1,034
A30	1,566	-6,898	1,468	-5,817

A continuación se muestra la figura que compara el HNR de la segunda etapa, es decir, del filtrado de Kalman:

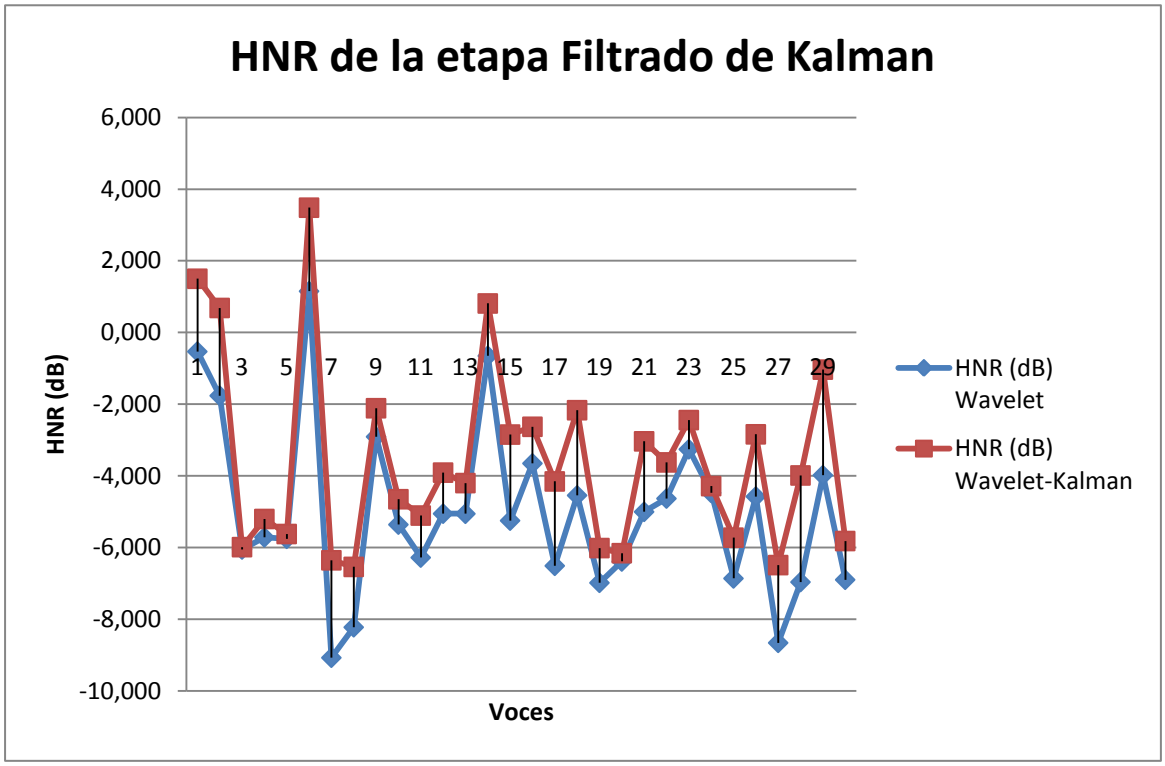

Figura 5.11: Comparación del HNR en la segunda etapa.

Como se puede apreciar claramente, el HNR de las voces después de ser procesadas por la segunda etapa utilizando la técnica de Filtrado de Kalman (en rojo, Wavelet-Kalman) es mayor que el HNR de las voces de la primera etapa (en azul, Wavelet).

Realizando el análisis estadístico, se debe comentar que los datos del HNR en esta etapa cumplen el criterio de normalidad, es decir, podemos asumir la normalidad de los datos. De cara a conocer la normalidad de los datos se han realizado dos pruebas: test de Kolmogorov-Smirnov [Fasano+87] y el de Shapiro-Wilk [Shapiro65]. Ambas pruebas han dado como resultado que la distribución de los datos del HNR para la voz esofágica sí está normalizada y la significancia de los datos es mayor que el 5%, lo cual indica que se pueden asumir la normalidad de los datos en esta etapa. Por lo tanto, podemos utilizar la prueba estadística del "T-student" [Park+11].

Una vez que podemos asumir la normalidad de los datos, se realiza la prueba "T-student" [Park+11] para comparar los datos. Esta prueba nos muestra que los datos que no son iguales: los datos anteriores y posteriores al procesado de la segunda etapa. La significancia es menor a un 5%, concretamente, $p<0,0001$. Esto nos indica que se rechaza la hipótesis nula, en la que se dice que las medias de los datos anterior y posterior a la segunda etapa son iguales.

Esto junto con la Figura 5.11 presentada anteriormente nos muestra que en esta etapa se produce una mejora del parámetro HNR.

Si nos fijamos en el Shimmer, se puede observar que esta etapa no experimenta un cambio considerable. De hecho, produce un aumento medio de dicho parámetro de un 0,014 dB. Algo que podemos considerar poco significativo. Podemos observar estos resultados en la siguiente figura:

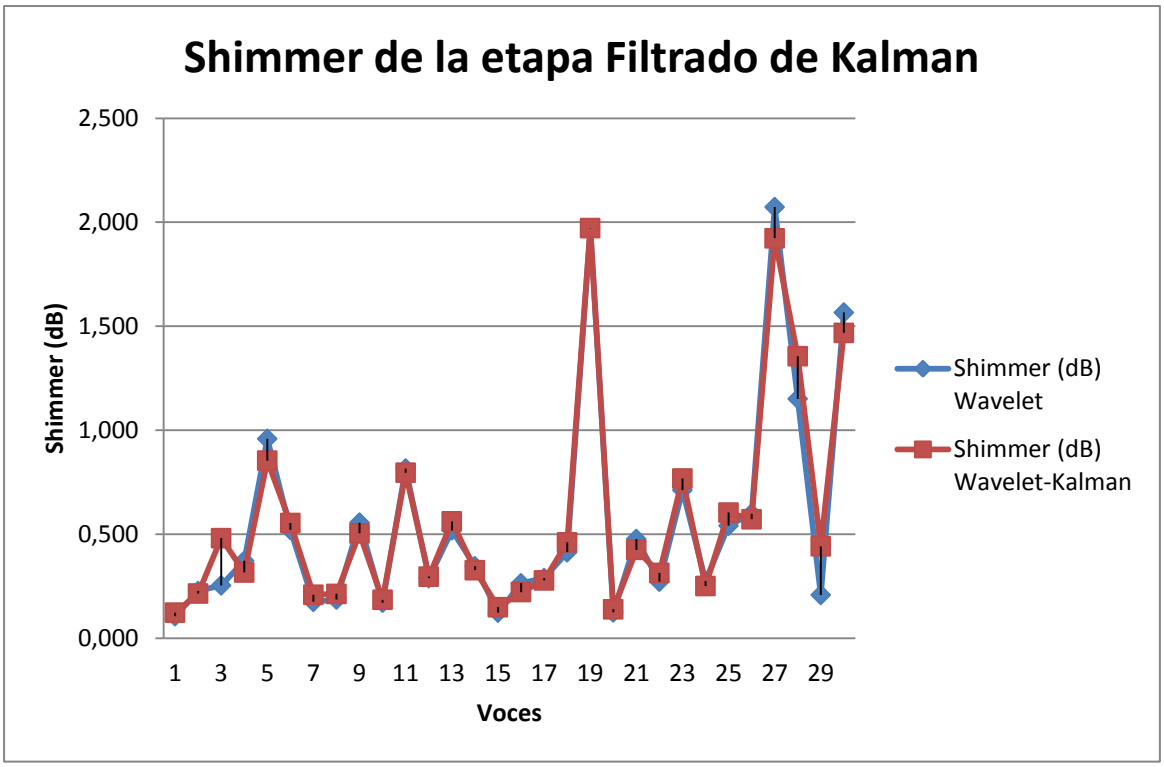

Figura 5.12: Comparación del Shimmer en la segunda etapa

Como se puede apreciar claramente, el Shimmer de las voces después de ser procesadas por la segunda etapa utilizando la técnica del Filtrado de Kalman (en rojo, Wavelet-Kalman) es similar al shimmer de las voces de la salida de la primera etapa (en azul, Wavelet).

Si analizamos los datos desde un punto de vista estadístico, se debe comentar que los datos del shimmer no cumplen el criterio de normalidad para la segunda etapa, es decir, no podemos asumir la normalidad en los datos. De cara a conocer la normalidad de los datos se han realizado dos pruebas: test de Kolmogorov-Smirnov [Fasano+87] y el de Shapiro-Wilk [Shapiro65]. Ambas pruebas han dado como resultado que la distribución de los datos del shimmer que se tienen en cuenta en esta etapa no están normalizadas. La significancia de los datos es menor que 0,001 lo cual indica que no se pueden asumir la normalidad de los datos. Una vez que no podemos asumir la normalidad de los datos, se realiza la prueba Wilcoxon [Wilcoxon45] para comparar los datos.

Esta prueba nos muestra que los datos del shimmer anteriores y posteriores a la segunda etapa puede que tengan medias iguales. La significancia de la prueba es p=0,565 y, por tanto, se no se puede rechazar la hipótesis nula que dice: "La mediana entre los datos anteriores y posteriores a la etapa filtrado de Kalman son iguales". Este resultado nos indica que no se pueden extraer conclusiones en el parámetro shimmer.

5.2.1.3 Pruebas de la Etapa de Estabilización de Polos

En esta etapa, se recogen las voces procesadas en la etapa de filtrado de Kalman y se procesan en esta tercera etapa mediante la estabilización de los polos. En esta etapa también, el parámetro en el que se hace hincapié es el HNR.

En la Tabla 5.4 se meustran los valores obtenidos en la etapa anterior (etapa de filtrado de Kalman) con los obtenidos en esta etapa tanto para el Shimmer como para el HNR.

Tabla 5.4: Resultados de la etapa Estabilización de Polos

	Procesadas (Kalman)		Procesadas (Polos)	
	Shimmer (dB)	**HNR (dB)**	**Shimmer (dB)**	**HNR (dB)**
A1	0,123	1,501	0,101	2,009
A2	0,215	0,686	0,186	1,534
A3	0,481	-5,987	0,454	-5,602
A4	0,316	-5,202	0,367	-4,542
A5	0,854	-5,621	0,886	-5,895
A6	0,554	3,484	0,513	4,173
A7	0,209	-6,354	0,184	-3,695
A8	0,213	-6,539	0,194	-3,708
A9	0,504	-2,111	0,435	-1,392
A10	0,185	-4,650	0,180	-4,046
A11	0,796	-5,106	0,775	-3,033
A12	0,296	-3,908	0,265	-3,173
A13	0,562	-4,200	0,543	-4,032
A14	0,327	0,812	0,318	3,203
A15	0,149	-2,846	0,133	-0,818
A16	0,222	-2,629	0,219	-1,147
A17	0,278	-4,154	0,263	-3,489
A18	0,461	-2,166	0,398	-0,140
A19	1,972	-6,013	1,527	-6,092
A20	0,139	-6,156	0,130	-6,095
A21	0,425	-3,035	0,336	-2,804
A22	0,314	-3,623	0,263	-3,081
A23	0,768	-2,442	0,654	-2,348
A24	0,251	-4,284	0,209	-4,017
A25	0,604	-5,723	0,518	-5,088
A26	0,572	-2,839	0,561	-2,711
A27	1,923	-6,488	1,412	-6,293
A28	1,356	-3,987	0,968	-3,832
A29	0,442	-1,034	0,324	0,612
A30	1,468	-5,817	1,355	-5,288

Como se puede observar en la tabla, los valores que mejores resultados presentan con respecto al HNR son los etiquetados como A7, A8, A14-A16, A18 y A29 ya que éstos son los que mayor incremento presentan en decibelios. La mejora media presentada en esta etapa en el HNR es de 0,853 dB. Esto se puede apreciar en la siguiente figura:

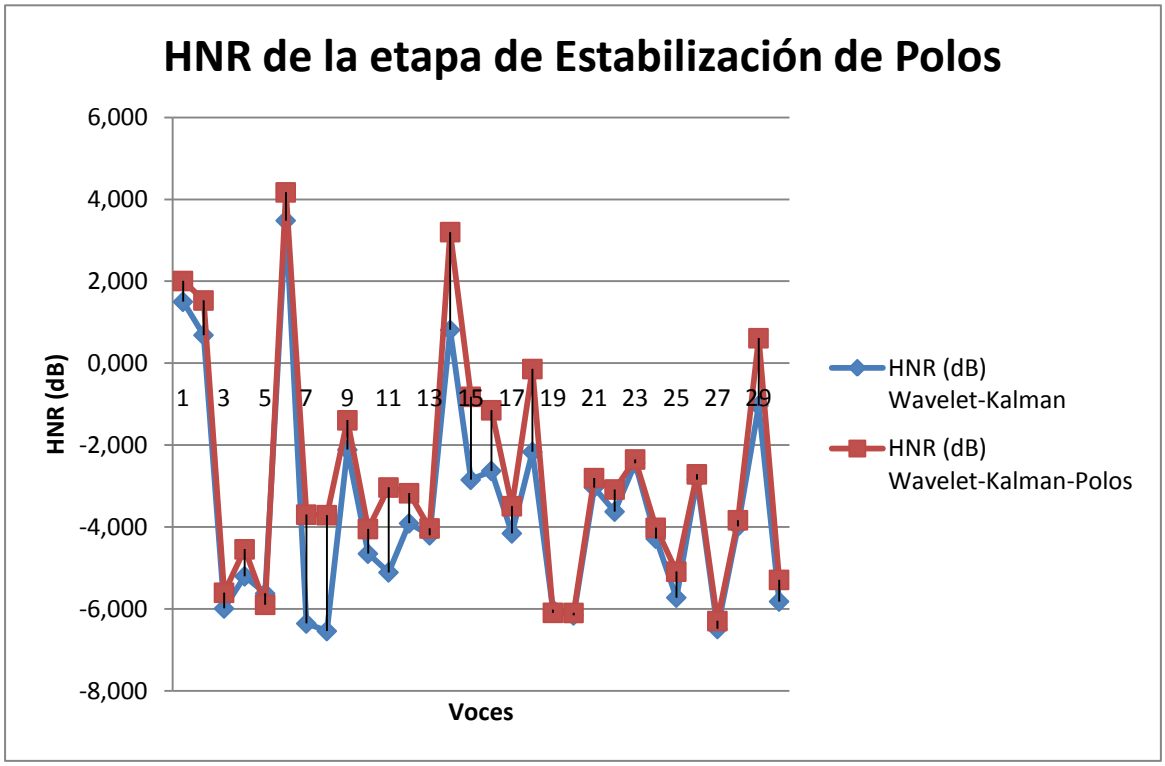

Figura 5.13: Comparación del HNR en la tercera etapa.

Como se puede apreciar claramente, el HNR de las voces después de ser procesadas por la segunda etapa utilizando la técnica de Estabilización de polos (en rojo, Wavelet-Kalman-Polos) es mayor que el HNR de las voces de la segunda etapa (en azul, Wavelet-Kalman). Esto supone una mejora clara de este parámetro, pero para obtener una mayor certeza realizamos a continuación el análisis estadístico.

En esta ocasión los datos del HNR en esta etapa no cumplen el criterio de normalidad, es decir, no podemos asumir la normalidad de los datos. De cara a conocer la normalidad de los datos se han realizado dos pruebas: test de Kolmogorov-Smirnov [Fasano+87] y el de Shapiro-Wilk [Shapiro65]. Ambas

pruebas han dado como resultado que la distribución de los datos del HNR para la voz esofágica no está normalizada y la significancia de los datos es menor que el 5%, lo cual indica que no se pueden asumir la normalidad de los datos en esta etapa. Por lo tanto, utilizaremos la prueba estadística Wilcoxon [Wilcoxon45].

Una vez realizada la prueba de cara a comparar los datos, la prueba nos muestra que los datos que no son iguales: los datos anteriores y posteriores al procesado de la tercera etapa. La significancia es menor a un 5%, concretamente, $p<0,0001$. Esto nos indica que se rechaza la hipótesis nula, en la que se dice que las medias de los datos anterior y posterior a la tercera etapa son iguales. Esto junto con la Figura 5.13 presentada anteriormente nos muestra que en esta etapa se produce una mejora del parámetro HNR.

Si nos fijamos en el Shimmer, se puede observar que esta etapa no experimenta un cambio considerable. De hecho, produce una disminución media de dicho parámetro de un 0,077 dB. Algo que podemos considerar poco significativo. Podemos observar estos resultados en la siguiente figura:

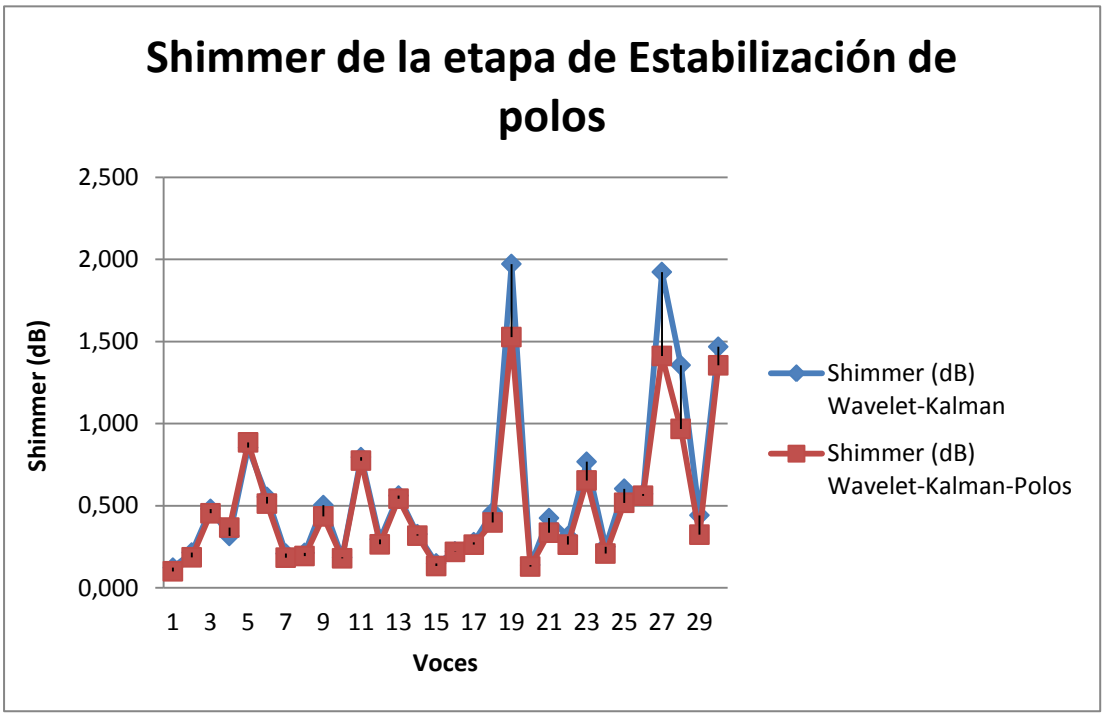

Figura 5.14: Comparación del Shimmer en la tercera etapa

Como se puede apreciar claramente, el Shimmer de las voces después de ser procesadas por la tercera etapa utilizando la técnica de Estabilización de polos (en rojo, Wavelet-Kalman-Polos) es similar al shimmer de las voces de la salida de la segunda etapa (en azul, Wavelet-Kalman), exceptuando las voces A19 y A27 que sí presentan una mejora significativa. Esto indica que hay una mejora aunque sea pequeña.

Si analizamos los datos desde un punto de vista estadístico, se debe comentar que los datos del shimmer no cumplen el criterio de normalidad para la tercera etapa, es decir, no podemos asumir la normalidad en los datos. De cara a conocer la normalidad de los datos se han realizado dos pruebas: test de Kolmogorov-Smirnov [Fasano+87] y el de Shapiro-Wilk [Shapiro65]. Ambas pruebas han dado como resultado que la distribución de los datos del shimmer que se tienen en cuenta en esta etapa no están normalizadas. La significancia de los datos es menor que 0,001 lo cual indica que no se pueden asumir la normalidad de los datos. Una vez que no podemos asumir la normalidad de los datos, se realiza la prueba Wilcoxon [Wilcoxon45] para comparar los datos.

Esta prueba nos muestra que los datos del shimmer anteriores y posteriores a la tercera etapa puede que tengan medias iguales. La significancia de la prueba es $p<0,001$ y, por tanto, se se puede rechazar la hipótesis nula que dice: "La mediana entre los datos anteriores y posteriores a la etapa de estabilización de polos son iguales". Este resultado nos indica que a pesar de que la mejora en el parámetro shimmer es pequeña, existe dicha mejora.

5.2.1.4 Análisis Global de las Tres Etapas

Es interesante mostrar la mejora de los parámetros Shimmer y HNR de todas las etapas. Esto nos muestra cuál es la mejora de ambos parámetros y gráficamente se puede observar las diferencias entre las etapas.

A continuación se muestra la gráfica de la evolución del shimmer:

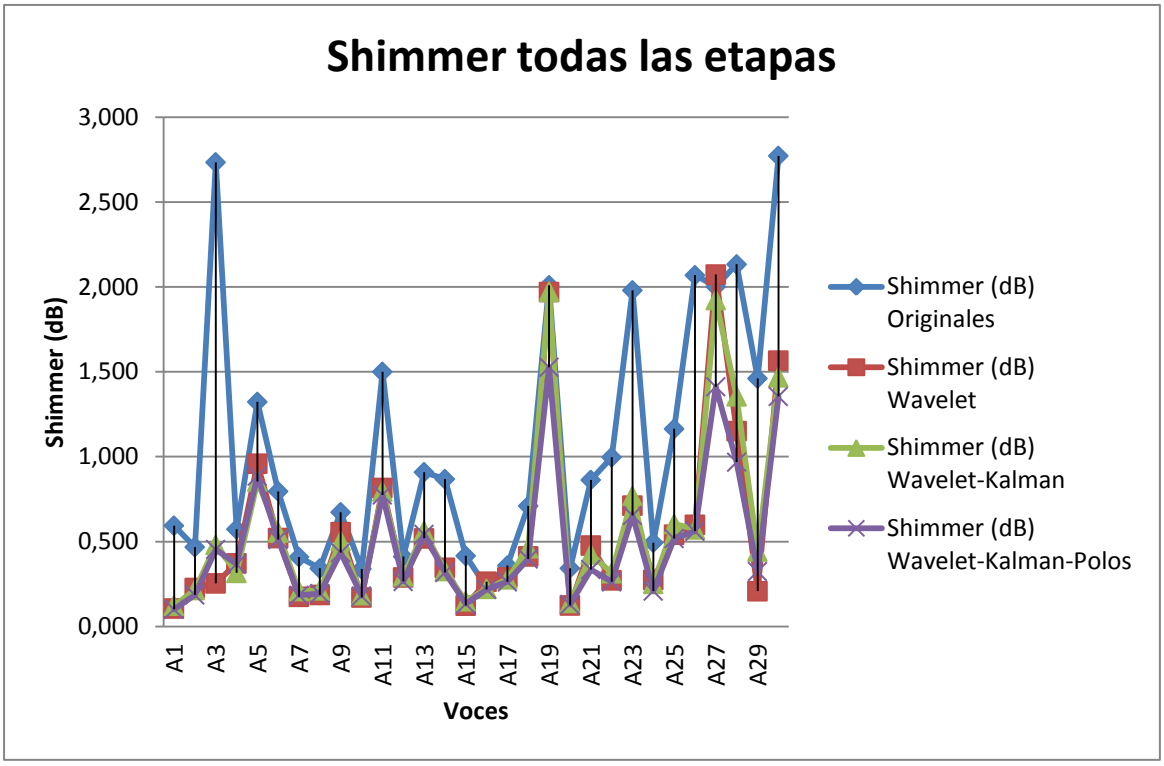

Figura 5.15: Comparación del shimmer de todas las etapas

Como podemos observar claramente, el shimmer experimenta una mejora global teniendo en cuenta todas las etapas ya que como muestra la figura, el shimmer después de todo el procesado (morado, Wavelet-Kalman-Polos) es mucho menor que el shimmer original (Azul, originales). Concretamente, se produce una disminución media global de 0,576 dB lo cual supone una mejora significativa con respecto a la original.

Las voces que experimentan una mayor mejora son las etiquetadas como A3, A23, A26 y A29-A30.

Desde un punto de vista estadístico se corroboran estos resultados. Para el shimmer, tal y como se ha reflejado en cada una de las etapas y esta ocasión también, no se puede asumir la normalidad de los mismos. Recordamos que en esta ocasión se tienen en cuenta para esta prueba los datos de entrada, las voces esofágicas originales y la salida de la etapa de "Mejora de la calidad de la voz esofágica". Se han realizado de nuevo las pruebas de Kolmogorov-Smirnov

[Fasano+87] y el de Shapiro-Wilk [Shapiro65] y ambas nos muestran que los datos no presentan normalidad con una significancia menor al 5%.

Se realiza, por tanto, la prueba de Wilcoxon [Wilcoxon45]. Esta prueba nos muestra que los datos del shimmer original y los datos procesados tras la etapa de "Mejora de la Calidad de la Voz Esofágica" no son iguales. La significancia de la prueba es menor que 0,001 ($p<0,0001$) y, por tanto, se rechaza la hipótesis nula que dice: "La mediana entre los datos originales y los datos obtenidos tras el procesado son iguales". Este resultado junto con la Figura 5.15 evidencia que existe una mejora significativa en el parámetro shimmer.

Resulta interesante comparar todos los datos al mismo tiempo. Para ello, se realiza el test de Friedman [Friedman39] que efectúa el análisis por parejas por rangos de muestras relacionadas en este caso. Esta prueba muestra una significancia menor que 0,001 ($p<0,0001$) y, por tanto, se rechaza la hipótesis nula que dice: "La distribuciones de las varianzas de las distintas muestras son las mismas".

Tabla 5.5: Comparación de las distintas muestras por parejas en el Shimmer

Comparación de muestras para el Shimmer	**Significancia**
Originales – Etapa 1 (Originales vs. Wavelets)	$p<0,0001$
Originales – Etapa1 + Etapa 2 (Originales vs. Wavelets - Kalman)	$p<0,0001$
Originales – Etapa1 + Etapa 2+ Etapa 3 (Originales vs. Wavelets - Kalman - Polos)	$p<0,0001$
Etapa1 - Etapa1 + Etapa 2 (Wavelets vs. Wavelets - Kalman)	$p = 0,424$
Etapa1 - Etapa1 + Etapa 2+ Etapa 3 (Wavelets vs. Wavelets - Kalman - Polos)	$p = 0,007$
Etapa2 - Etapa1 + Etapa 2+ Etapa 3 (Kalman vs. Wavelets - Kalman - Polos)	$p<0,0001$

Esta tabla nos muestra claramente atendiendo a las tres primeras filas que las distribuciones de las distintas etapas con respecto a la original son totalmente

diferentes, con lo que se vuelva corroborar que se produce una mejora en la calidad de las voz.

Para visualizar esta mejora de la calidad de la voz en las diferentes muestras, es ilustrador mostrar la media de cada una de las etapas:

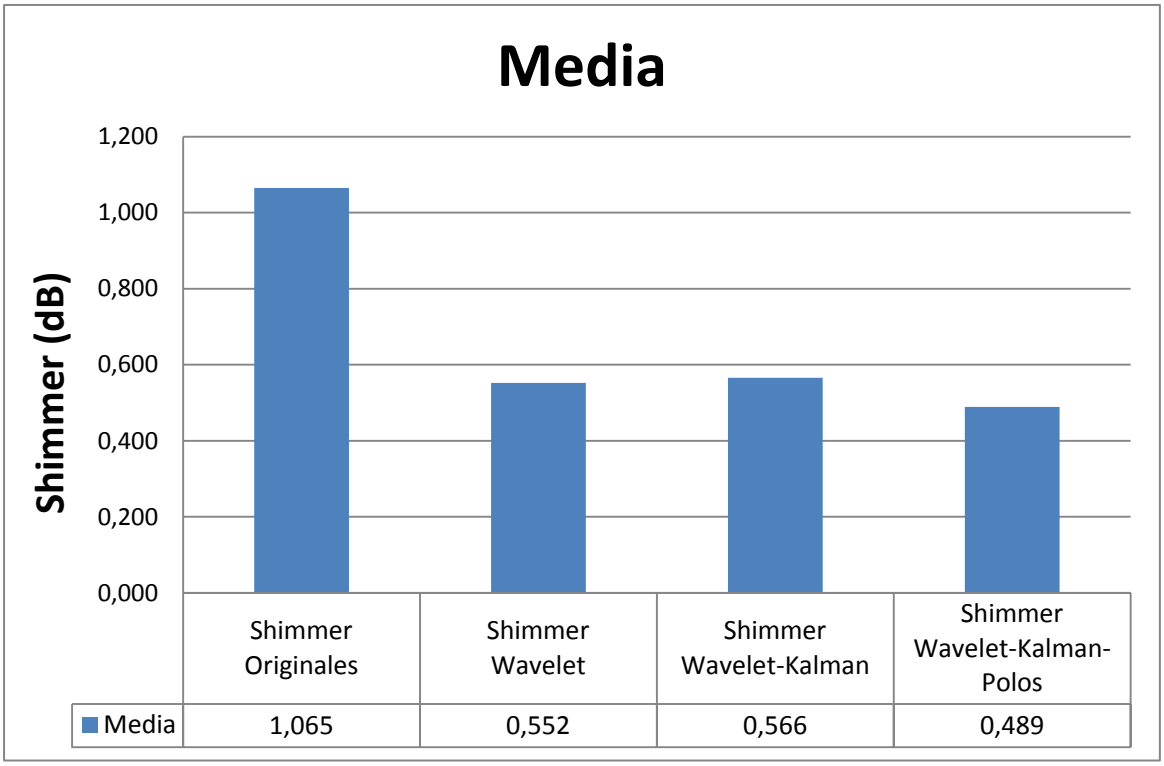

Figura 5.16: Media del Shimmer de las distintas etapas

El test de Friedman [Friedman39], Tabla 5.5, junto con la media del Shimmer en las distintas etapas del algoritmo, Figura 5.16, nos dice que las distribuciones de las muestras para el parámetro Shimmer son diferentes y, que la media de las muestras es menor en cualquiera de las etapas si la comparamos con la original. Con lo que se concluye que se produce una mejora de la calidad de la voz ya que se constata una disminución del Shimmer.

Realizando algo similar con el HNR obtenemos los siguientes resultados comparando el HNR en las distintas etapas:

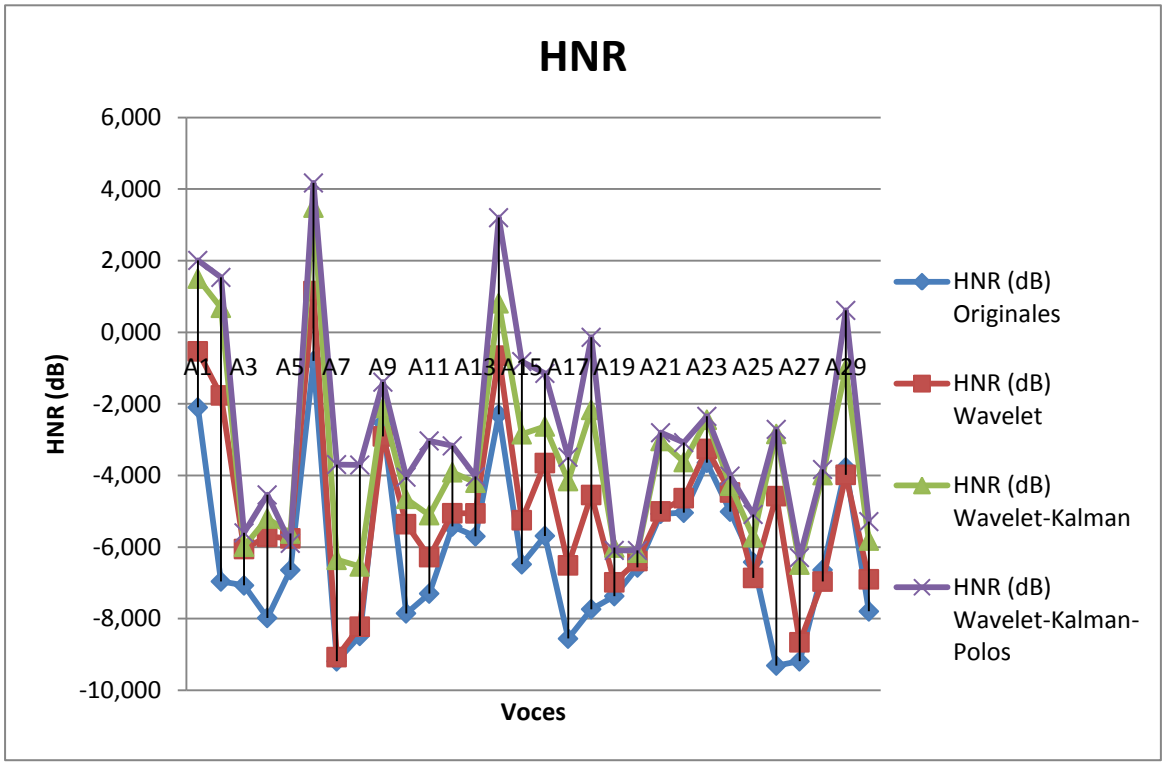

Figura 5.17: Comparación del HNR de todas las etapas

Como podemos observar claramente, el HNR experimenta una mejora global teniendo en cuenta todas las etapas ya que como muestra la figura, el HNR después de todo el procesado (morado, Wavelet-Kalman-Polos) es mucho mayor que el HNR original (Azul, originales). Concretamente, se produce un aumento medio global de 3,459 dB lo cual supone una mejora significativa con respecto a la original.

Las voces que experimentan una mayor mejora son las etiquetadas como A2, A14-A15, A17, A18, A26 y A29.

Para confirmar la validez de los resultados obtenidos se analizan estadísticamente. En esta ocasión, no se pueden asumir la normalidad de los datos. Se han realizado de nuevo las pruebas de Kolmogorov-Smirnov [Fasano+87] y el de Shapiro-Wilk [Shapiro65] y ambas nos muestran que los datos no presentan normalidad con una significancia menor al 5%.

Se realiza, por tanto, la prueba de Wilcoxon [Wilcoxon45]. Esta prueba nos muestra que los datos del HNR original y los datos procesados tras la etapa de "Mejora de la Calidad de la Voz Esofágica" no son iguales. La significancia de la prueba es menor que 0,001 ($p<0,0001$) y, por tanto, se rechaza la hipótesis nula que dice: "La mediana entre los datos originales y los datos obtenidos tras el procesado son iguales". Este resultado junto con la Figura 5.17 evidencia que existe una mejora significativa en el parámetro HNR.

Al igual que para el shimmer, comparamos todos los datos al mismo tiempo para el HNR realizando para ello el test de Friedman [Friedman39]. Esta prueba muestra una significancia menor que 0,001 ($p<0,0001$) en la mayoría de los casos y, por tanto, se rechaza la hipótesis nula que dice: "La distribuciones de las varianzas de las distintas muestras son las mismas".

Tabla 5.6: Comparación de las distintas muestras por parejas en el HNR

Comparación de muestras para el HNR	**Significancia**
Originales – Etapa 1 (Originales vs. Wavelets)	$p = 0,029$
Originales – Etapa1 + Etapa 2 (Originales vs. Wavelets - Kalman)	$p<0,0001$
Originales – Etapa1 + Etapa 2+ Etapa 3 (Originales vs. Wavelets - Kalman - Polos)	$p<0,0001$
Etapa1 - Etapa1 + Etapa 2 (Wavelets vs. Wavelets - Kalman)	$p = 0,003$
Etapa1 - Etapa1 + Etapa 2+ Etapa 3 (Wavelets vs. Wavelets - Kalman - Polos)	$p<0,0001$
Etapa2 - Etapa1 + Etapa 2+ Etapa 3 (Kalman vs. Wavelets - Kalman - Polos)	$p = 0,075$

Esta tabla nos muestra claramente atendiendo a las tres primeras filas que las distribuciones de las distintas etapas con respecto a la original son totalmente diferentes, con lo que se vuelva corroborar que se produce una mejora en la calidad de las voz.

Para visualizar esta mejora de la calidad de la voz en las diferentes muestras, es ilustrador mostrar la media de cada una de las etapas:

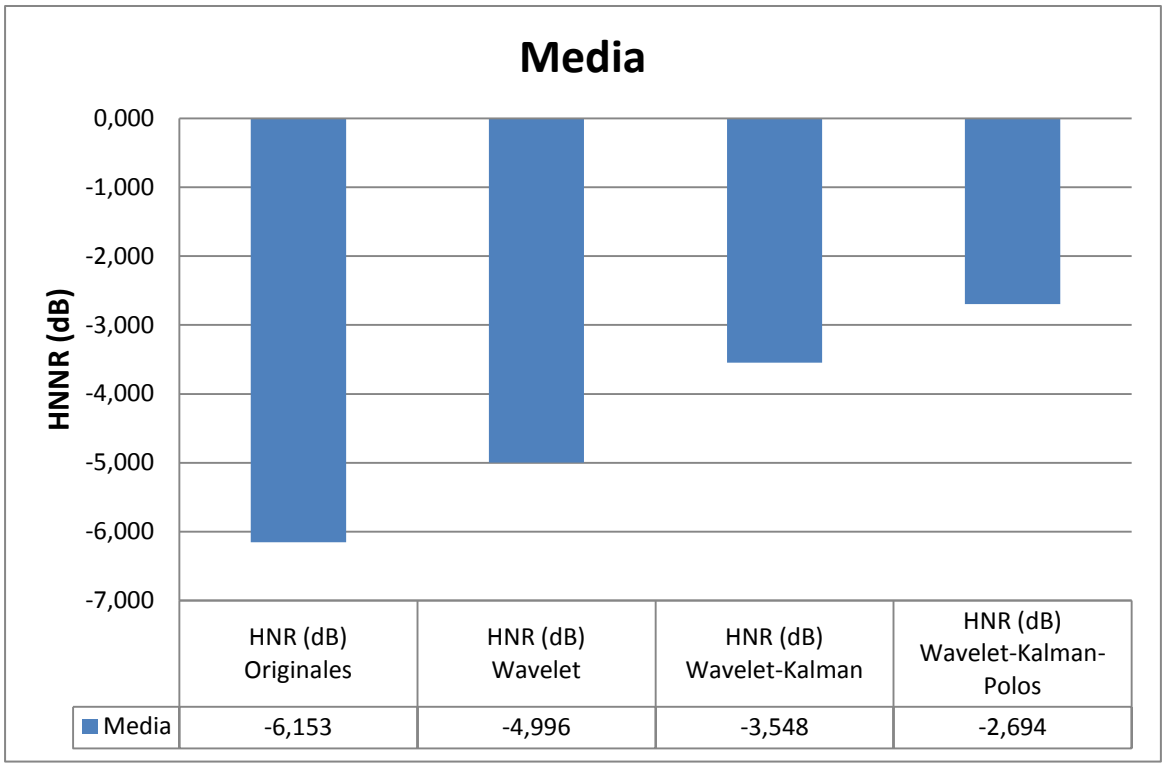

Figura 5.18: Media del HNR de las distintas etapas

El test de Friedman [Friedman39], Tabla 5.6, junto con la media del HNR en las distintas etapas del algoritmo, Figura 5.18, nos dice que las distribuciones de las muestras para el parámetro HNR son diferentes y, que la media de las muestras es menor en cualquiera de las etapas si la comparamos con la original. Con lo que se concluye que se produce una mejora de la calidad de la voz ya que se constata un aumento del HNR.

5.2.1.5 Valoración subjetiva de la mejora de la voz esofágica

La calidad de la voz esofágica también ha sido evaluada de forma subjetiva mediante la prueba de Mean Opinion Square (MOS) [Murphy+07b]. Los 30 fonemas originales de la base de datos y las correspondientes voces procesadas han sido presentados a 30 personas oyentes.

La escala aceptabilidad de la voz en la prueba MOS comprende los valores entre 1 (insatisfactorio) y 5 (excelente).

En la tabla que se muestra a continuación se presentan los valores obtenidos en la prueba MOS para los distintos tipos de voces, concretamente: voces originales, voces procesadas con la etapa wavelet, voces procesadas con el filtro de Kalman con todos los ruidos utilizados y las voces procesadas con las tres etapas, es decir, con el algoritmo propuesto.

Tabla 5.7: Evaluación de la voz de forma subjetiva

Evaluación de la voz de forma subjetiva	**MOS**
Voz esofágica original	1,74
Voz procesada con la etapa Wavelet (Bior 6.8)	2,03
Wavelet + Kalman con ruido blanco	2,11
Wavelet+ Kalman con ruido marrón	2,39
Wavelet+ Kalman con ruido esofágico en los momentos de silencio	2,47
Wavelet + Kalman con ruido rosa	1,99
Wavelet +Kalman con ruido violeta	2,45
Wavelet +Kalman (esofágico) + Estabilización de polos	3,05

Subjetivamente, una vez procesadas las voces mediante el algoritmo propuesto reduce sustancialmente el ruido de aspiración del esófago, sin aparente reducción de la inteligibilidad.

5.2.2 Pruebas de la "Mejora de la Parametrización de la Voz Esofágica"

El algoritmo de parametrización de la voz esofágica está basado en un procedimiento iterativo que extrae los instantes de pitch o las marcas de la señal de voz para calcular el pitch. Para estas pruebas se han utilizado 10 voces esofágicas y diez voces sanas.

En el capítulo 4, diseño del algoritmo, se muestran 7 bloques (B1,…, B7) que describen el algoritmo de parametrización de la voz. Los bloques B1 y B5, el algoritmo base, describen la misma sub-etapa aunque con una asignación de parámetros diferente y da muestra del carácter iterativo del algoritmo. Tal y como se ha descrito en dicho apartado, los bloques B2, B3 y B4 corresponden a la clasificación y a la asignación de parámetros del algoritmo. Obviamente, estos bloques no generan ningún resultado con lo que a continuación se mostrarán los resultados de los otros bloques.

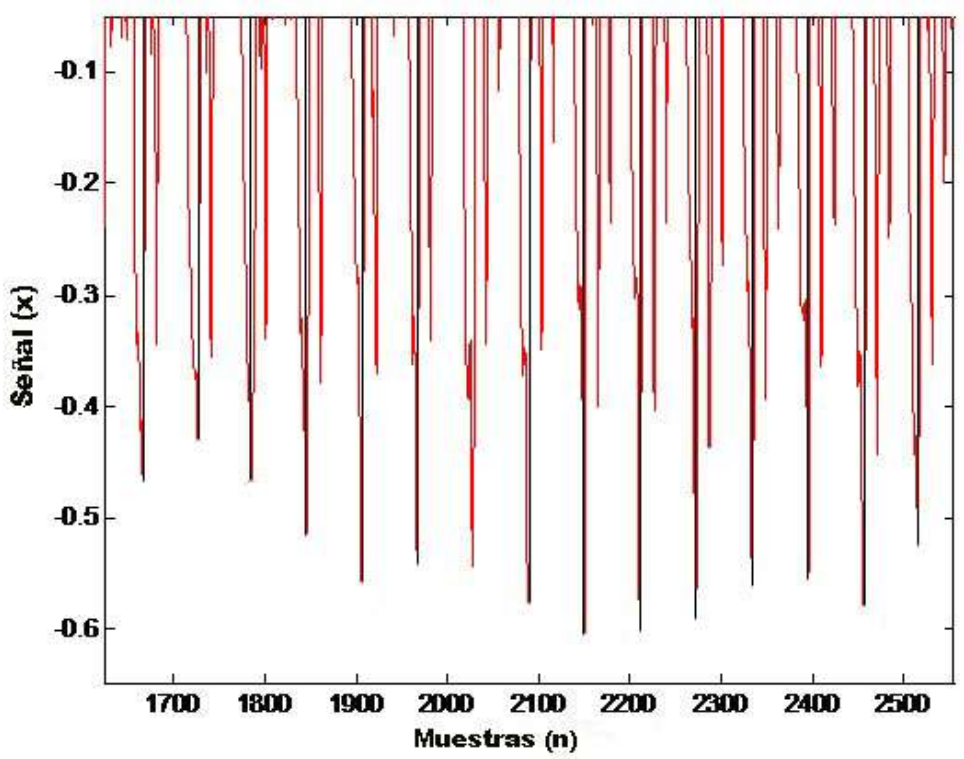

Figura 5.19: Picos negativos de la señal

5.2.2.1 Pruebas de la etapa del Algoritmo Base

El primer paso de este algoritmo es quedarse con los picos negativos de la señal que se muestra en la Figura 5.19.

Posteriormente se realiza la transformada rápida de Fourier de cara a llevar cabo el cálculo de la sonoridad con los parámetros asignados en este módulo y descritos en el capítulo 4. Si se superpone el resultado del paso de la sonoridad a la señal original de la voz, se puede observar que los picos de la señal es donde mayor energía hay (Figura 5.20).

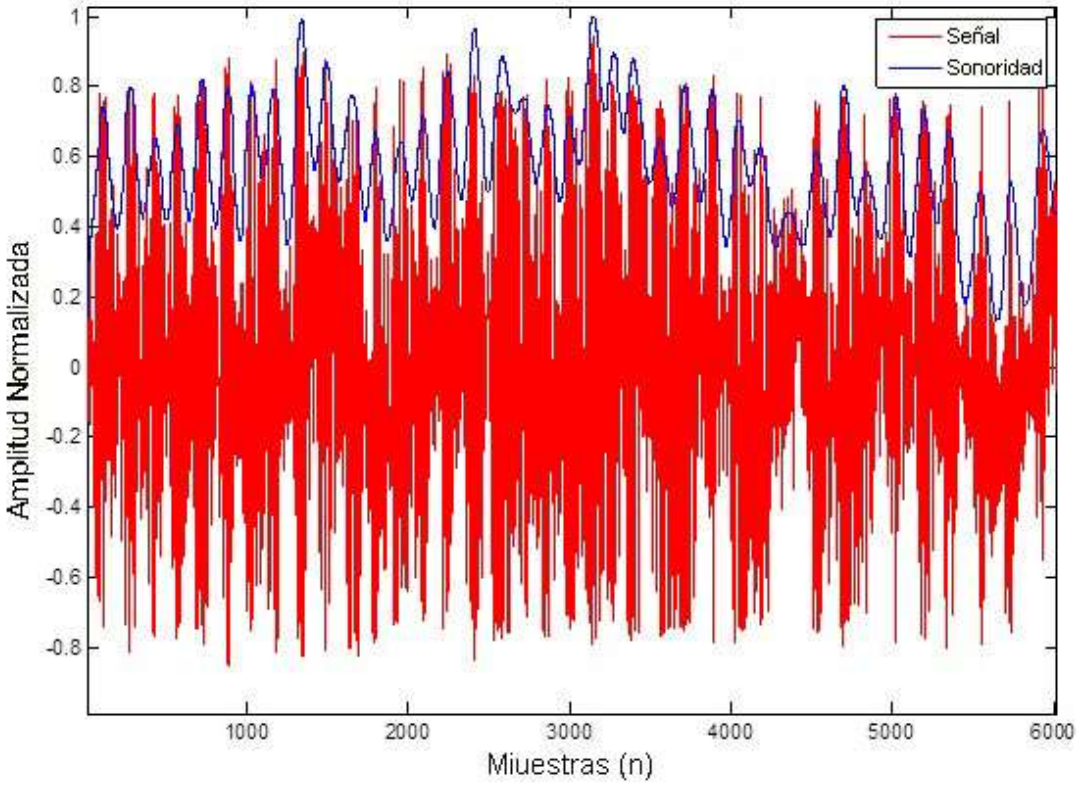

Figura 5.20: La señal de la voz y sonoridad de la señal

El siguiente paso es obtener los instantes de pitch en aquellos lugares donde la sonoridad es mayor. Un resultado tipo de una señal de voz de este proceso se muestra en la Figura 5.21. A continuación se ajustan los instantes de pitch a los mínimos de la señal cuya figura se ha mostrado en el capítulo anterior (Figura 4.39).

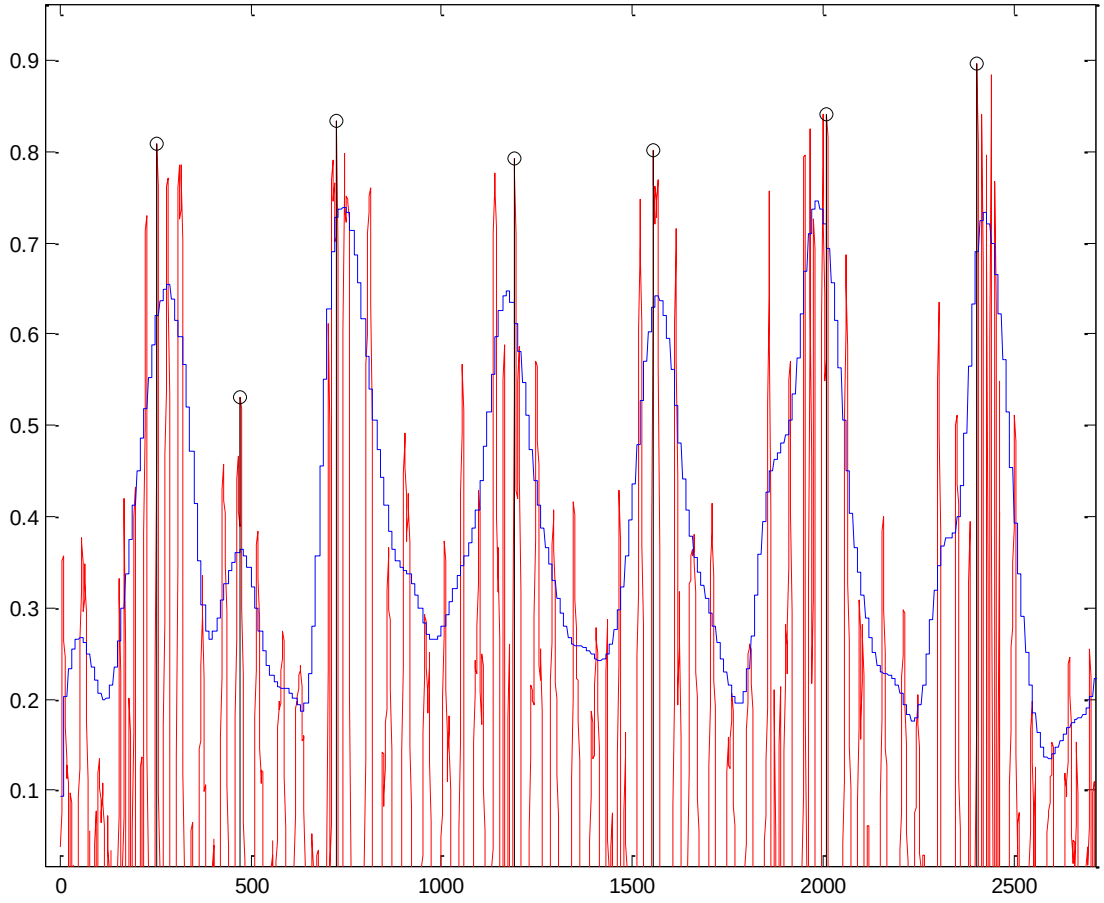

Figura 5.21: Señal de voz, sonoridad e instantes de pitch

Con todos estos resultados se obtiene un vector de unos y ceros donde los unos representan a los instantes de pitch y las demás posiciones se rellenan con ceros. Una vez que tenemos este vector es inmediato obtener el pitch de la señal. Recordamos en este punto el carácter iterativo del algoritmo ya que esta etapa se realiza dos veces para un mayor refinamiento de dicho algoritmo (sub-etapas B1 y B5). La única diferencia radica en la asignación de parámetros de este algoritmo.

5.2.2.2 Pruebas de la etapa de las Acciones Correctoras

Esta etapa se encarga de los picos que no se detectan. A pesar de que el algoritmo base es muy efectivo en la búsqueda de instantes de pitch, para una mayor precisión del algoritmo se realizan una comprobación de qué picos faltan o sobran en nuestro vector base.

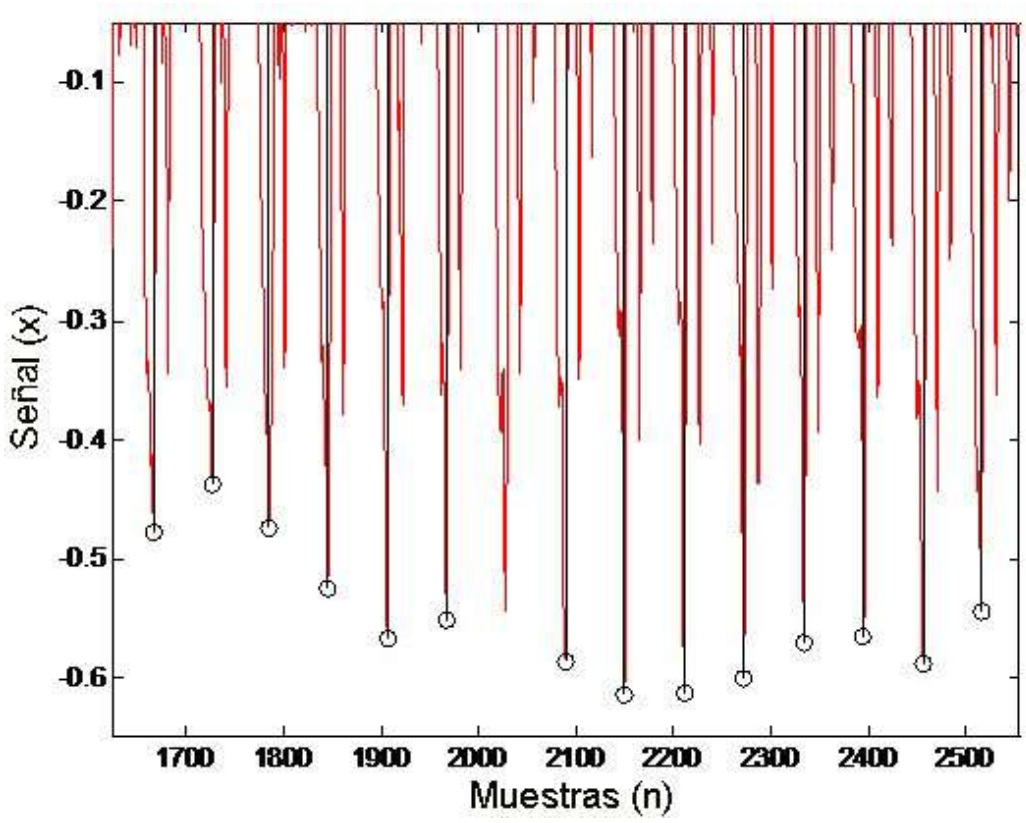

Figura 5.22: Picos no detectados

En la Figura 5.22 se muestra cómo existen picos que no son detectados en el algoritmo base y que justifican esta sub-etapa.

5.2.2.3 Pruebas del Cálculo de Parámetros

En esta sección se muestra todos los resultados obtenidos con el algoritmo propuesto y son comparados con los obtenidos por el MDVP (el Gold Standard). Con este mismo programa se han obtenido los datos a los que consideramos reales y, se obtienen estableciendo los instantes de pitch a mano. Se han realizado las medidas del pitch, jitter, shimmer y HNR para 10 voces esofágicas y 10 voces sanas con el fin de comparar cómo se comporta el algoritmo propuesto frente al Gold Standard.

En definitiva, tenemos dos medidores y queremos compararlos. Para ello, tenemos que ver la concordancia de las medidas [Toschke+08] realizadas con respecto a la medida real. Cabe señalar en este punto que no sería correcto realizar la correlación de ambas medidas ya que lo que se quiere conocer es la

concordancia entre las medidas y las reales [Sprent+00], es decir, distancias entre ellas.

5.2.2.3.1 Pruebas de las medidas del pitch

A continuación, se muestras los resultados de las medidas realizadas para el pitch para voces sanas.

Tabla 5.8: Medidas de pitch para voces sanas

Voces Sanas	Medidas de Pitch (Hz)		
	Algoritmo	**MDVP**	**Reales**
A1	215,435	214,736	215,422
A2	110,670	110,692	110,657
A3	132,243	132,275	132,231
A4	102,705	107,286	102,696
A5	112,998	113,077	112,987
A6	120,185	120,086	120,183
A7	109,595	109,647	109,589
A8	114,117	114,249	114,109
A9	120,185	120,086	120,183
A10	193,125	193,139	193,085

La primera columna representan las etiquetas de las voces sanas para el fonema "a". La segunda columna nos muestra las medidas del pitch obtenidas con el algoritmo propuesto en este trabajo de investigación. En la tercera columna figuran las medidas del Multi-Dimensional Voice Program [Deliyeski93] y, finalmente, la cuarta columna está constituida por lo que se consideran las medidas reales. Éstas son medidas con el mismo programa mencionado, el Gold standard, pero asignando los instantes de pitch a mano. En los datos reflejados en la segunda columna, los medidos con el programa MDVP, se establecen las marcas de pitch automáticamente y como podremos comprobar más adelante no lo realiza correctamente con las voces esofágicas.

Cabe mencionar que las medidas de las voces sanas tienen valores mayores que 100 Hz en todos los casos, cosa que no sucede en las voces esofágicas cuyos valores rondan los 60 Hz como se muestra en la Tabla 5.10.

Se puede observar atendiendo a los resultados que ambas medidas son bastante similares a la medida real. No obstante, posteriormente realizaremos un análisis estadístico para la obtención de conclusiones más decisivas.

De cara a realizar el estudio estadístico de estos datos se considera que, por un lado, se debe reflejar la concordancia entre los datos medidos con el algoritmo propuesto y los datos reales y, por otro lado, la concordancia de los datos obtenidos con el MDVP y los reales. Para ello, se realizará la gráfica de Bland-Altman [Altman+83] en la que se reflejan en el eje de las coordenadas las diferencias entre la medida a testear (la del algoritmo propuesto y la del MDVP) y la real, y en el eje de las abscisas la media de ambas (la medida a testear y la real). Es decir:

$$S(x,y) = S\left(\frac{a+b}{2}, (a-b)\right) \tag{5.1}$$

El gráfico es una figura de dispersión de las medidas y nos muestra las diferencias de las medidas frente a la media de las mismas. Si una medida más cerca del eje de las abscisas significa que la medida es buena ya que la distancia entre la medida "a" y la medida "b" es pequeña. El grado de distancia pequeña se da cuando se muestra la distancia en comparación con la media de ambas medidas.

La Figura 5.23 muestra la gráfica de Bland-Altman para las medidas de pitch de las voces sanas. Los datos en azul de la Figura 5.23 muestran las diferencias entre las medidas obtenidas con el algoritmo propuesto y las reales (eje de coordenadas), frente a la media de ambas (eje de las abscisas). Por otro lado, los datos en rojo muestran las diferencias entre los datos obtenidos con el MDVP (eje de coordenadas) y los reales, frente a la media de ambas (eje de las abscisas).

Se puede apreciar que ambas técnicas, el algoritmo y el MDVP, obtienen valores de pitch muy similares a las reales ya que la mayoría de sus puntos en la gráfica son cercanos al eje de las abscisas. No obstante, se puede apreciar cómo dos de los datos realizados con el MDVP se alejan de dicho eje arrojando un peor resultado que el algoritmo propuesto.

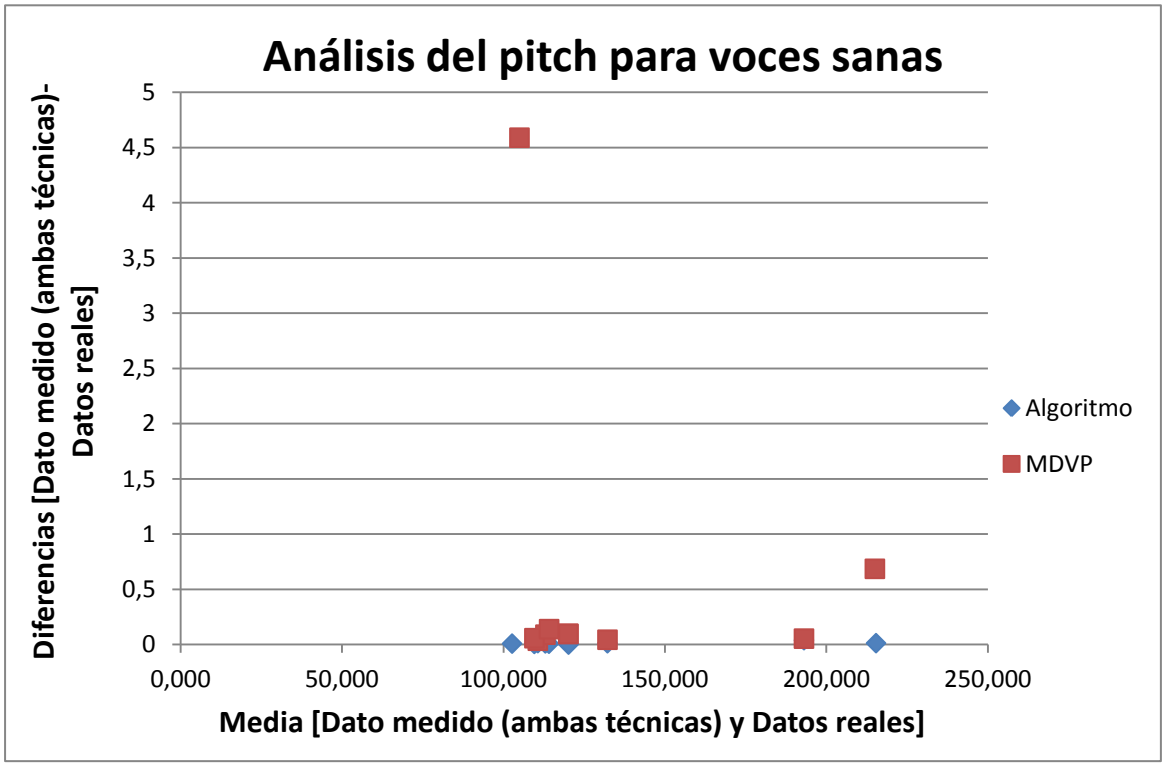

Figura 5.23: Bland-Altman para el pitch de voces sanas

Para obtener un mayor rigor estadístico, se ha realizado la prueba T-student [Park+11] o la prueba Wilcoxon [Wilcoxon45], según la normalidad de los datos. Para dicha prueba se han tenido en cuenta los datos de las diferencias entre la técnica testeada y las reales. Para una mayor claridad se muestran dichos datos en la Tabla 5.9.

De cara a realizar la prueba mencionada se ha hecho un análisis de la normalidad de los datos. En este caso todas las medidas en cuestión son normales con lo que se procede a realizar la prueba T-student. Ésta muestra un significancia para la voces sanas de p= 0,231. Esto nos dice que no se puede rechazar la hipótesis nula que dice que las medias de ambas muestras de datos son iguales. La significancia

muestra un 76,9% de diferencia entre las medias y se necesitaría un 95% para afirmar que las medias son totalmente diferentes y que se rechaza la hipótesis nula.

Tabla 5.9: Diferencias entre las técnicas y datos reales para las voces sanas

	Diferencias entre Voces Sanas	
	Algoritmo – Real (Hz)	**MDVP-Real (Hz)**
A1	0,013	0,686
A2	0,013	0,035
A3	0,012	0,044
A4	0,009	4,590
A5	0,011	0,09
A6	0,002	0,097
A7	0,006	0,058
A8	0,008	0,14
A9	0,002	0,097
A10	0,04	0,054

A pesar de que la media de la distancia entre el algoritmo propuesto y la muestra real (media de la distancia 0,0116 Hz) es menor que la distancia entre la del MDVP y el real (media de la distancia 0,5891 Hz), no se puede afirmar que de forma significativa sean distintas.

Como conclusión de este estudio se puede destacar que ambas técnicas dan buenos resultados para medir el pitch de las voces sanas y, que a pesar de que el algoritmo propuesto de media sea mejor que el del MDVP, no se puede decir que sea significativamente mejor.

En la Tabla 5.10 se muestran las medidas de pitch para las voces esofágicas. Como ya se ha mencionado, los valores del pitch para las voces esofágicas son menores. Se debe destacar que en una de las voces, en la etiquetada como A10 concretamente, el programa MDVP no ha sido capaz de dar un resultado ya que no ha podido establecer las marcas de pitch.

Tabla 5.10: Medidas de pitch para voces esofágicas

Voces Esofágicas	Medidas de Pitch (Hz)		
	Algoritmo	**MDVP**	**Reales**
A1	69,150	78,583	68,878
A2	60,249	65,978	62,004
A3	60,635	69,132	62,774
A4	84,281	82,508	83,319
A5	70,494	77,922	69,95
A6	61,495	77,996	61,379
A7	64,087	76,805	64,186
A8	61,448	79,991	61,193
A9	58,754	71,664	58,288
A10	101,448	ND[1]	104,587

La Figura 5.24 muestra la gráfica de Bland-Altman para las medidas de pitch de las voces sanas. Al igual que en el caso anterior, los datos en azul muestran las diferencias entre las medidas obtenidas con el algoritmo propuesto y las reales (eje de coordenadas), frente a la media de ambas (eje de las abscisas). Por otro lado, los datos en rojo muestran las diferencias entre los datos obtenidos con el MDVP (eje de coordenadas) y los reales, frente a la media de ambas (eje de las abscisas). En las gráficas de Bland-Altman que se mostrarán para las medidas del jitter, shimmer y HNR se seguirá esta misma elección para los colores, es decir, azul para el algoritmo propuesto y rojo para el MDVP.

Un simple vistazo de la Figura 5.24 nos indica que las distancias entre las medidas del MDVP y las reales están más alejadas del eje de las abscisas que las del algoritmo propuesto en este trabajo. Esto indica que el algoritmo propuesto tiene una distancia menor que el MDVP a las medidas reales y, por lo tanto, es un mejor medidor. Este dato deberá ser corroborado con el análisis estadístico que se mostrará más adelante.

[1] ND: No disponible. Para esta voz el programa MDVP no es capaz de mostrar ningún resultado.

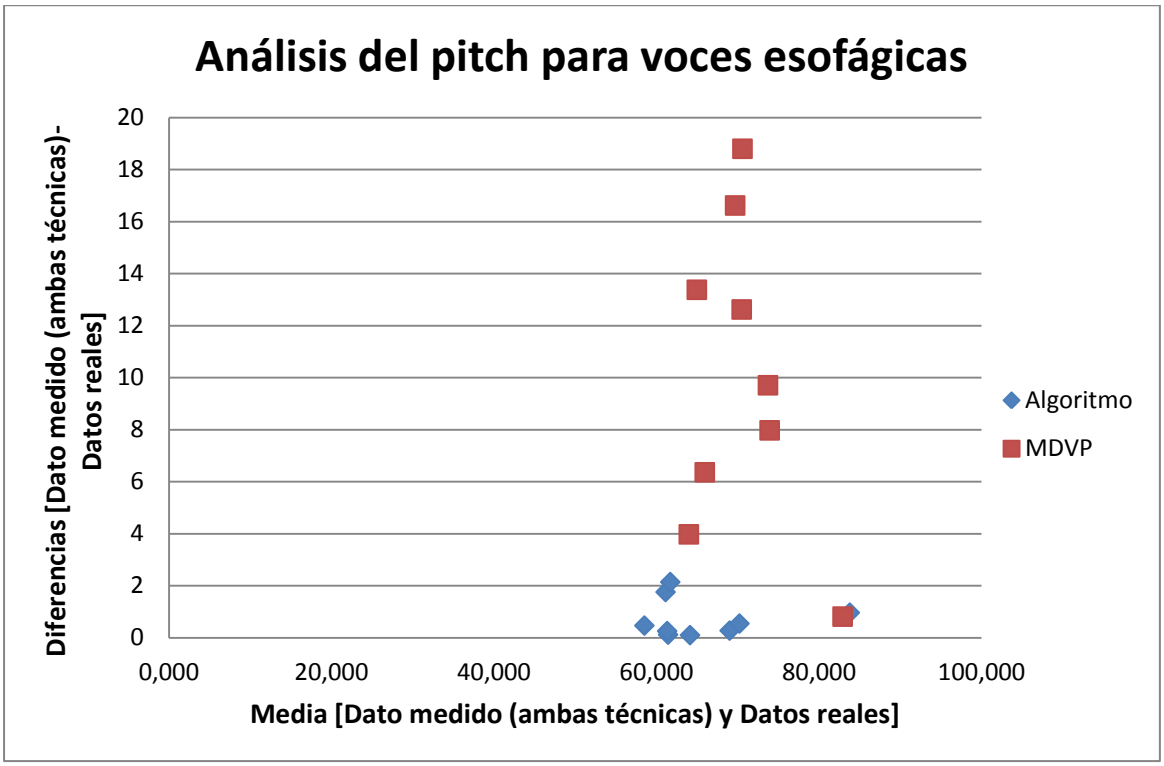

Figura 5.24: Bland-Altman para el pitch de voces esofágicas

Los datos de las distancias entre ambas técnicas y las reales se muestran en la Tabla 5.11 para una mayor claridad. En los estudios del jitter, shimmer y HNR no se mostrarán las diferencias ya que el procedimiento es el mismo.

Tabla 5.11: Diferencias entre las técnicas y datos reales para voces esofágicas

	Diferencias entre Voces Esofágicas	
	Algoritmo – Real (Hz)	**MDVP-Real (Hz)**
A1	0,272	9,705
A2	1,755	3,974
A3	2,139	6,358
A4	0,962	0,811
A5	0,544	7,972
A6	0,116	16,617
A7	0,099	12,619
A8	0,255	18,798
A9	0,466	13,376
A10	3,139	104,587

Se debe mencionar que la voz etiquetada como A10 para la columna del MDVP se ha tomado como cero el valor que el propio programa no ha conseguido medir. Este dato no será tenido en cuenta a la hora de realizar el análisis estadístico.

De cara a realizar el estudio estadístico, al igual que para las voces sanas, se puede asumir la normalidad de los datos con lo que se procede a realizar la prueba T-student. Dicha prueba arroja un resultado de $p = 0{,}002 < 0{,}05$ (95%), lo cual indica que se puede rechazar la hipótesis nula que dice que las medias de las distancias entre los datos reales y la de ambas técnicas son iguales. Es decir, las distancias entre ambas técnicas son significativamente diferentes con una probabilidad del 99,8%.

Por lo tanto, debido a que la media de las distancias entre el algoritmo propuesto y los datos reales (media=0,7342 Hz) es significativamente diferente a la media de las distancias entre el MDVP y los datos reales (media=10,0255 Hz), y atendiendo a los resultados de la Figura 5.24, se puede concluir que el algoritmo propuesto en este trabajo es significativamente mejor medidor que el MDVP para medir el pitch en las voces esofágicas.

5.2.2.3.2 Pruebas de las medidas del jitter

Realicemos a continuación para el jitter el mismo análisis que se ha realizado para el pitch. La Tabla 5.12 muestra las medidas obtenidas para el jitter en las diez voces sanas. De un simple vistazo se puede ver que las medidas obtenidas con el algoritmo propuesto se ajustan más que las obtenidas con el MDVP, percepción que se deberá corroborar con el análisis estadístico.

Se calculan de nuevo las distancias del algoritmo propuesto y las del MDVP con respecto a las medidas reales. Con esos resultados y teniendo en cuenta las medias se realiza la gráficas de Bland-Altman que se muestra en la Figura 5.25.

Tabla 5.12: Medidas del jitter para voces sanas

Voces Sanas	Medidas del Jitter (%)		
	Algoritmo	MDVP	Reales
A1	0,556	0,344	0,556
A2	0,576	0,326	0,576
A3	0,589	0,318	0,539
A4	0,915	1,328	0,915
A5	0,353	0,237	0,336
A6	0,313	0,153	0,298
A7	0,365	0,175	0,365
A8	0,528	0,432	0,453
A9	0,313	0,153	0,298
A10	1,299	0,713	1,299

Al igual que con el pitch los puntos azules corresponden a las distancias entre el algoritmo propuesto y las medidas reales, mientras que los puntos rojos son las diferencias entre las medidas del MDVP y las medidas reales.

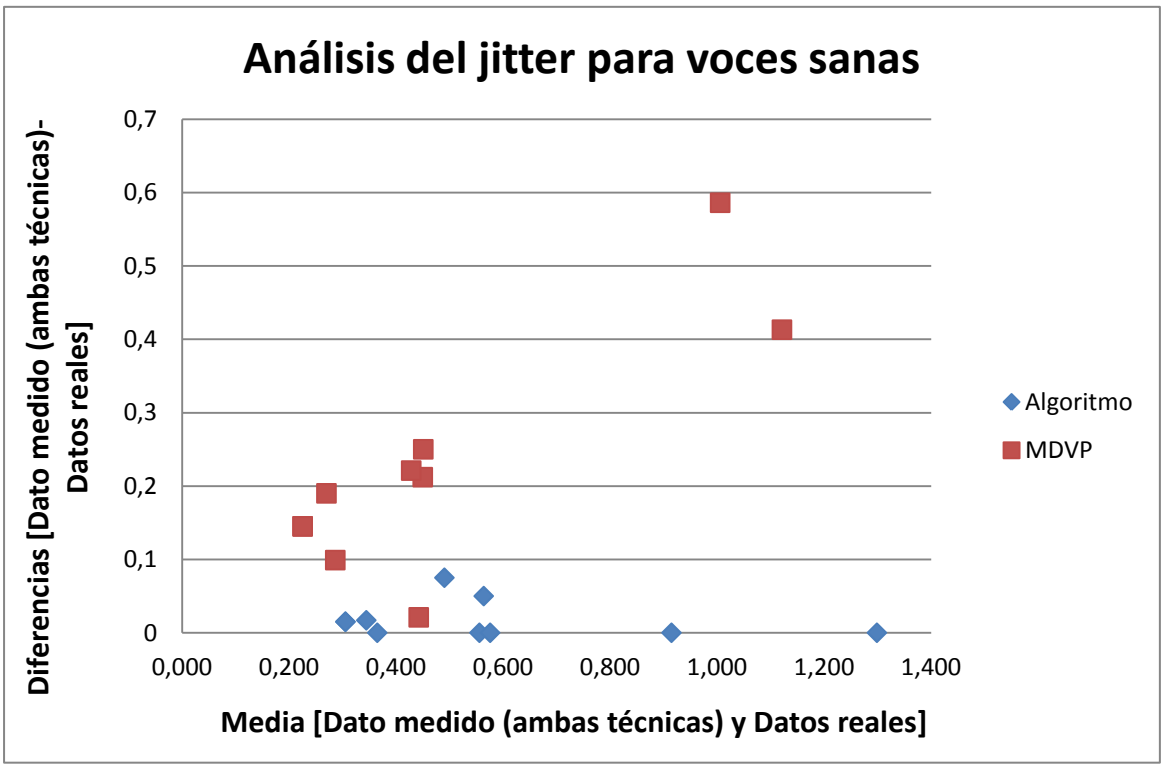

Figura 5.25: Bland-Altman para el jitter de voces sanas

Como se puede apreciar con gran claridad en la Figura 5.25, los puntos más cercanos al eje de las abscisas son los puntos del algoritmo propuesto, incluso algunos están sobre el mismo eje de las abscisas (distancia cero). Eso quiere decir que ciertas medidas coinciden exactamente con las medidas reales.

Para hacer un análisis estadístico más riguroso, tomamos de nuevo las distancias de ambas técnicas con respecto a las reales y miraremos a ver si sus medias son iguales. En esta ocasión, no se puede asumir la normalidad de los datos con lo que el estudio a realizar es la prueba Wilcoxon [Wilcoxon45]. Dicha prueba refleja un resultado de p = 0,007. Esto implica que se rechaza la hipótesis nula que dice que las medias de ambas muestras de datos son iguales. Se rechaza esta hipótesis al 99,3%(>95%) con lo que se puede afirmar que las medias son significativamente diferentes. La media de las diferencias entre el algoritmo y las mediadas reales es de 0,0172% y, la media de las distancias entre el MDVP y las medidas reales es de 0,2282%. Este resultado junto con la Figura 5.25 evidencia que el algoritmo propuesto es significativamente mejor medidor para la medida del jitter para voces sanas que el MDVP.

Tabla 5.13: Medidas del jitter para las voces esofágicas

Voces Esofágicas	**Medidas del Jitter (%)**		
	Algoritmo	**MDVP**	**Reales**
A1	3,587	2,503	3,587
A2	8,744	3,611	9,338
A3	5,785	3,067	4,599
A4	6,559	1,19	6,559
A5	5,442	2,958	5,442
A6	0,974	4,025	0,974
A7	7,462	2,965	5,312
A8	2,327	2,224	2,327
A9	4,391	4,61	1,848
A10	11,965	ND[2]	10,639

[2] ND: No disponible. Para esta voz el programa MDVP no es capaz de mostrar ningún resultado.

En la Tabla 5.13 se muestran los datos del jitter para las voces esofágicas. Al igual que sucede para el pitch, el programa MDVP no es capaz de dar un resultado para la voz etiquetada A10.

Calculando las diferencias y las medias para estos datos obtenemos la gráfica de Bland-Altman para el jitter de las voces esofágicas (Figura 5.26).

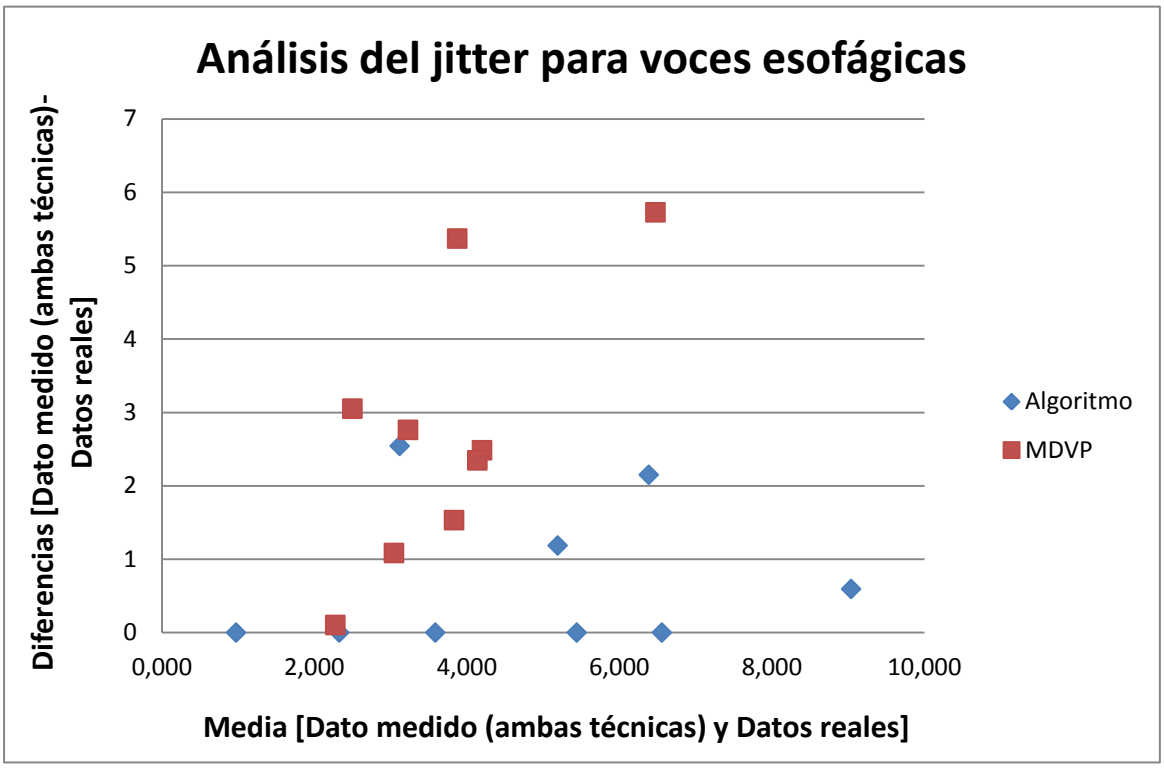

Figura 5.26: Bland-Altman para el jitter de voces esofágicas

Si nos fijamos en la Figura 5.26 parece que los puntos de las distancias entre el algoritmo propuesto y las reales (puntos azules) son más cercanos al eje de las abscisas lo cual indica que las diferencias a las medidas reales son menores que las del MDVP. Pero hay algunos puntos que tienen distancias parecidas.

Realizando el análisis estadístico de las diferencias, tal y como se ha hecho en los casos anteriores, comprobaremos si nuestra percepción es cierta. Tomando las diferencias de ambos medidores con respecto a las reales, observamos que no se puede asumir la normalidad de los datos. Por lo tanto, se realiza la prueba de Wilcoxon [Wilcoxon45]. Dicha prueba muestra una significancia de p=0,008. Esto nos indica que se rechaza la hipótesis nula y nos dice que las medias de las

distancias entre ambas técnicas y los datos reales son significativamente diferentes a un 99,8%%(>95%). La media de las diferencias entre el algoritmo y las mediadas reales es de 0,7192% y, la media de las distancias entre el MDVP y las medidas reales es de 2,7176%. Este resultado junto con la Figura 5.26 evidencia que el algoritmo propuesto es significativamente mejor medidor para la medida del jitter para voces sanas que el MDVP. Recordamos que también en esta ocasión no se han tenido en cuenta los datos no válidos que ha dado el MDVP.

Como conclusión en las medidas del jitter, se puede decir que el medidor del algoritmo propuesto es significativamente mejor que el medidor del MDVP.

5.2.2.3.3 Pruebas de las medidas del shimmer

Pasemos ahora al análisis de las medidas del shimmer. La Tabla 5.14 muestra las medidas del shimmer para las voces sanas. En esta ocasión no parece tan evidente que un medidor sea mejor que el otro a la vista de los datos.

Tabla 5.14: Medidas del shimmer para las voces sanas

Voces Sanas	Medidas del Shimmer (dB)		
	Algoritmo	**MDVP**	**Reales**
A1	0,431	0,269	0,593
A2	0,623	0,509	0,53
A3	0,823	0,847	1,887
A4	0,541	0,78	0,816
A5	0,338	0,301	0,314
A6	0,298	0,322	0,273
A7	0,383	0,474	0,378
A8	0,729	0,614	0,718
A9	0,298	0,322	0,272
A10	0,300	0,284	0,328

Realizando las diferencias y las medias, como para los otros parámetros, obtenemos la gráfica de Bland-Altman del shimmer (Figura 5.27).

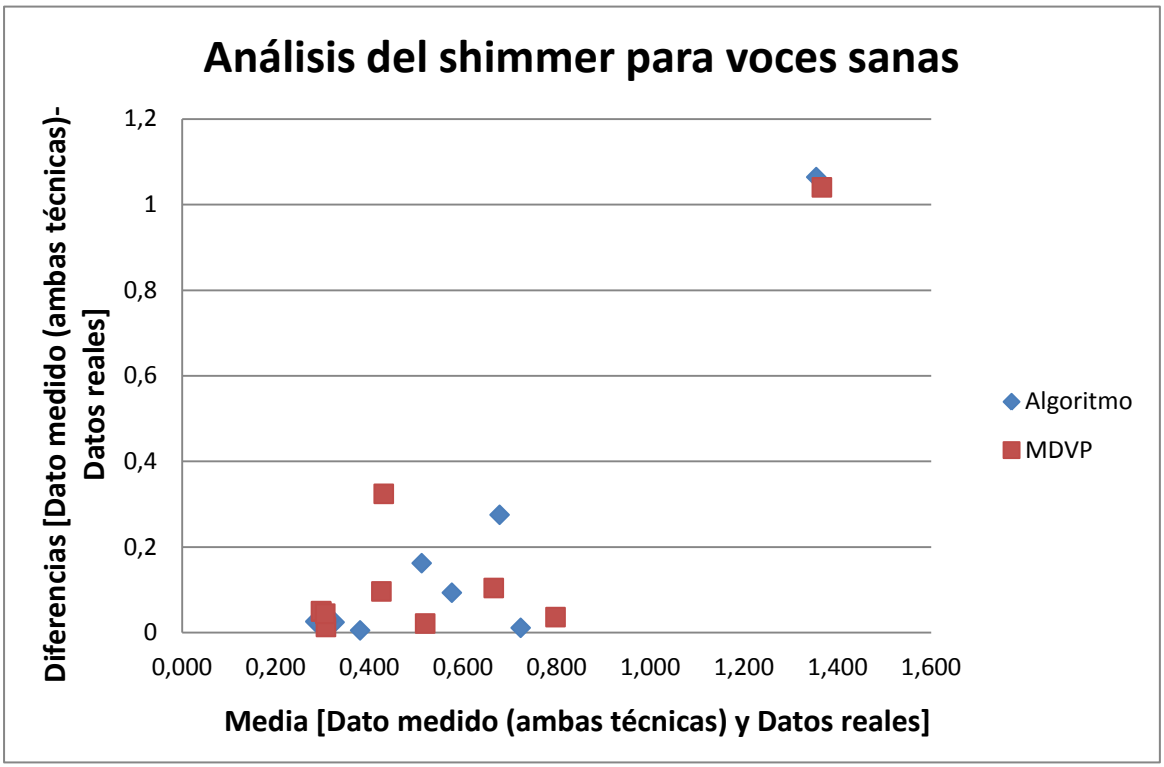

Figura 5.27: Bland-Altman para el shimmer de voces sanas

Como se puede apreciar en la Figura 5.27 todos los puntos tanto para un medidor como para el otro están bastante dispersos. No se puede apreciar con gran claridad que las distancias de un medidor sean menores que las del otro con respecto a las reales.

Debemos corroborar esta apreciación visual con un análisis estadístico. Se comparan las diferencias de los medidores con respecto a las medidas reales, tal y como se ha realizado con los otros parámetros. En esta ocasión no se puede asumir la normalidad de los datos con lo que se procede a realizar la prueba de Wilcoxon [Wilcoxon45]. La prueba muestra un resultado de p=0,507. Esto nos dice que no se puede, en absoluto, rechazar la hipótesis nula y que, por tanto, puede que las medias sean iguales. La media de las diferencias entre el algoritmo y las mediadas reales es de 0,1713 dB y, la media de las distancias entre el MDVP y las medidas reales es de 0,1777 dB. Este resultado junto con la Figura 5.27

muestra que no se puede decir que ninguno de los dos medidores sea mejor que el otro. Se puede decir que ambos medidores son igualmente "buenos" o "malos".

Veamos ahora los datos del shimmer de las voces esofágicas en la Tabla 5.15.

Tabla 5.15: Medidas del shimmer para las voces esofágicas

Voces Esofágicas	**Medidas del Shimmer (dB)**		
	Algoritmo	**MDVP**	**Reales**
A1	0,364	0,773	0,339
A2	1,319	0,900	1,500
A3	1,344	2,537	1,222
A4	0,381	0,351	0,412
A5	0,906	0,824	0,909
A6	0,816	1,941	0,868
A7	0,344	0,292	0,416
A8	0,328	0,382	0,359
A9	0,712	2,925	0,950
A10	0,334	ND[3]	0,342

Como en los anteriores parámetros, en la voz etiquetada A10 el medidor MDVP no da un resultado para el shimmer. Al igual que en los anteriores parámetros, en el caso de las voces esofágicas el algoritmo propuesto parece dar mejores resultados que el MDVP.

Realizando las diferencias y las medias, como para los otros parámetros, obtenemos la gráfica de Bland-Altman del shimmer para voces esofágicas (Figura 5.28). En esta ocasión se aprecia claramente que las diferencias entre el MDVP y los datos reales están más alejados del eje de las abscisas que las diferencias entre el algoritmo propuesto y los datos reales. Esto muestra que, en principio, el medidor con el algoritmo propuesto da mejores resultados que el del MDVP para

[3] ND: No disponible. Para esta voz el programa MDVP no es capaz de mostrar ningún resultado.

las voces esofágicas, pero esto deberemos corroborarlo con en el análisis estadístico.

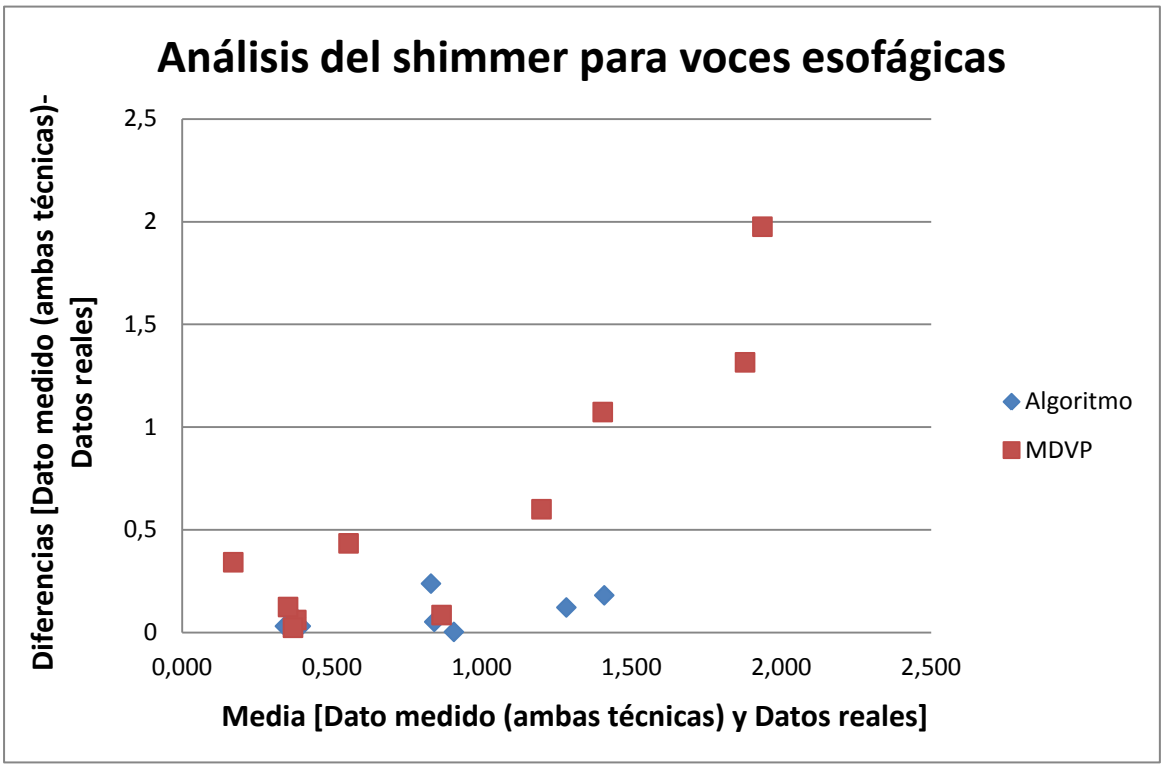

Figura 5.28: Bland-Altman para el shimmer de las voces esofágicas

De cara a realizar un análisis más riguroso, comparamos las medias de las diferencias entre cada uno de los medidores y los datos reales. En esta ocasión tampoco se puede asumir la normalidad de los datos y, por tanto, se realiza la prueba de Wilcoxon [Wilcoxon45]. Ésta muestra una significancia de p=0,011, lo cual supone que se rechaza la hipótesis nula que dice que las medias de las diferencias entre ambos medidores y las medidas reales son iguales. De hecho, son significativamente diferentes con un 98,9% (>95%). La media de las diferencias entre el algoritmo y las mediadas reales es de 0,0838 dB y, la media de las distancias entre el MDVP y las medidas reales es de 0,6322 dB. Este resultado junto con la Figura 5.28 evidencia que el algoritmo propuesto es significativamente mejor medidor para la medida del shimmer para voces esofágicas que el MDVP. Recordamos que también en esta ocasión no se han tenido en cuenta los datos no válidos que ha dado el MDVP.

5.2.2.3.4 Pruebas de las medidas del HNR

Finalmente, mostramos las medidas del HNR para las voces sanas (Tabla 5.16).

Tabla 5.16: Medidas del HNR para las voces sanas

Voces Sanas	Medidas del HNR (dB)		
	Algoritmo	MDVP	Reales
A1	-1,192	-1,210	-0,540
A2	3,420	3,413	3,500
A3	-0,236	-0,302	0,025
A4	6,980	6,854	7,538
A5	2,440	2,312	2,664
A6	1,520	1,123	1,370
A7	4,100	5,562	4,744
A8	9,400	9,523	8,267
A9	0,400	0,684	0,133
A10	6,330	6,435	6,077

Realizando las diferencias y las medias, como para los otros parámetros, obtenemos la gráfica de Bland-Altman del HNR (Figura 5.29).

Se puede apreciar que aunque los datos, en general, están bastante dispersos, las diferencias entre el algoritmo propuesto y las medidas reales están más cercanas al eje de las abscisas (puntos azules) que las diferencias con el MDVP (puntos en rojo). No obstante, como en anteriores ocasiones, se debe confirmar esta percepción con el análisis estadístico.

Para llevar a cabo el análisis estadístico, comparamos las medias de las diferencias entre cada uno de los medidores y los datos reales. En esta ocasión tampoco se puede asumir la normalidad de los datos y, por tanto, se realiza la prueba de Wilcoxon [Wilcoxon45]. Ésta muestra una significancia de p=0,005, lo cual supone que se rechaza la hipótesis nula que dice que las medias de las diferencias entre ambos medidores y las medidas reales son iguales. De hecho, son significativamente diferentes con un 99,5% (>95%).

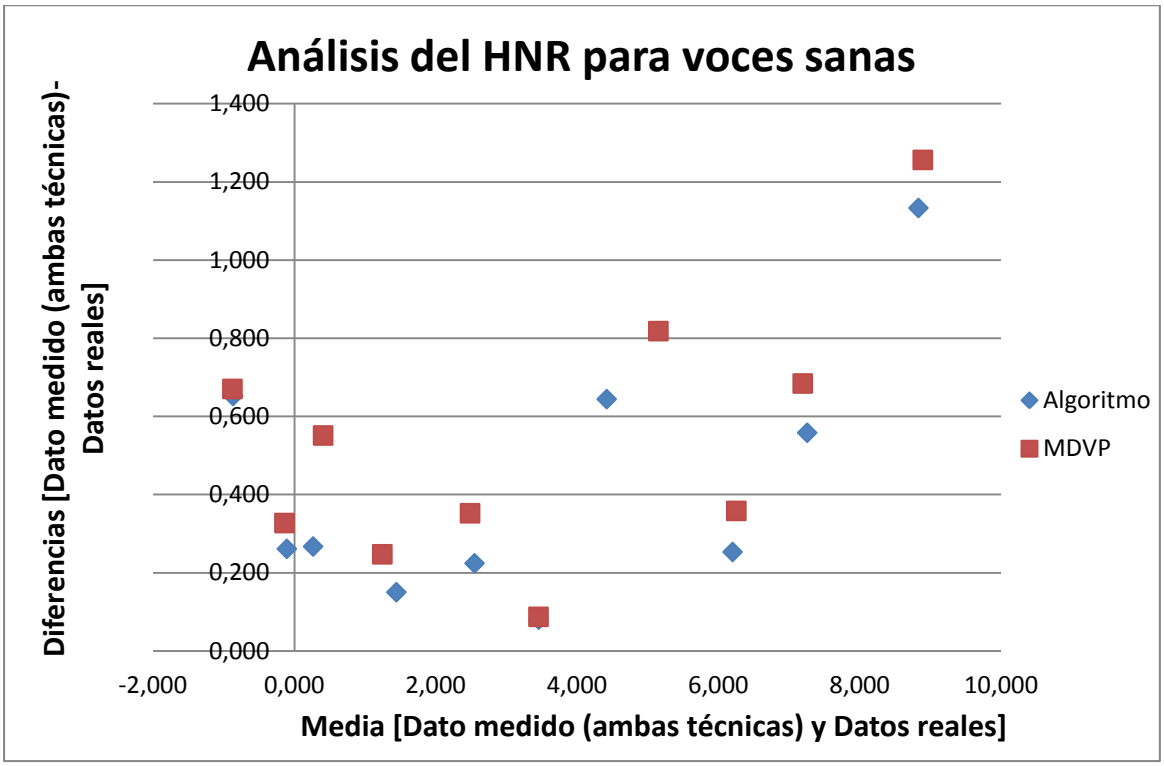

Figura 5.29: Bland-Altman para el HNR de voces sanas

La media de las diferencias entre el algoritmo y las medidas reales es de 0,422 dB y, la media de las distancias entre el MDVP y las medidas reales es de 0,535 dB. A pesar de que la diferencia entre las media no parece muy elevada, la prueba del Wilcoxon muestra que las medias son significativamente diferentes con un 99,8%. Este resultado junto con la Figura 5.29 evidencia que el algoritmo propuesto es significativamente mejor medidor para la medida del HNR para voces sanas que el MDVP.

Pasemos a analizar las medidas del HNR para las voces esofágicas (Tabla 5.17). Como en los anteriores parámetros, en la voz etiquetada A10 el medidor MDVP no da un resultado para el HNR. Al igual que en los anteriores parámetros, en el caso de las voces esofágicas el algoritmo propuesto parece dar mejores resultados que el MDVP.

Tabla 5.17: Medidas del HNR para las voces esofágicas

Voces Esofágicas	Medidas del HNR (dB)		
	Algoritmo	**MDVP**	**Reales**
A1	-8,998	-4,523	-8,900
A2	-8,381	-5,627	-8,340
A3	-7,867	-6,125	-7,780
A4	-7,299	-2,541	-7,400
A5	-8,465	-5,895	-8,460
A6	-4,833	-1,357	-5,340
A7	-5,972	-2,962	-6,120
A8	-2,394	-5,687	-2,590
A9	-6,479	-1,549	-6,520
A10	-8,668	ND[4]	-8,720

Realizando las diferencias y las medias, obtenemos la gráfica de Bland-Altman del HNR para voces esofágicas (Figura 5.30).

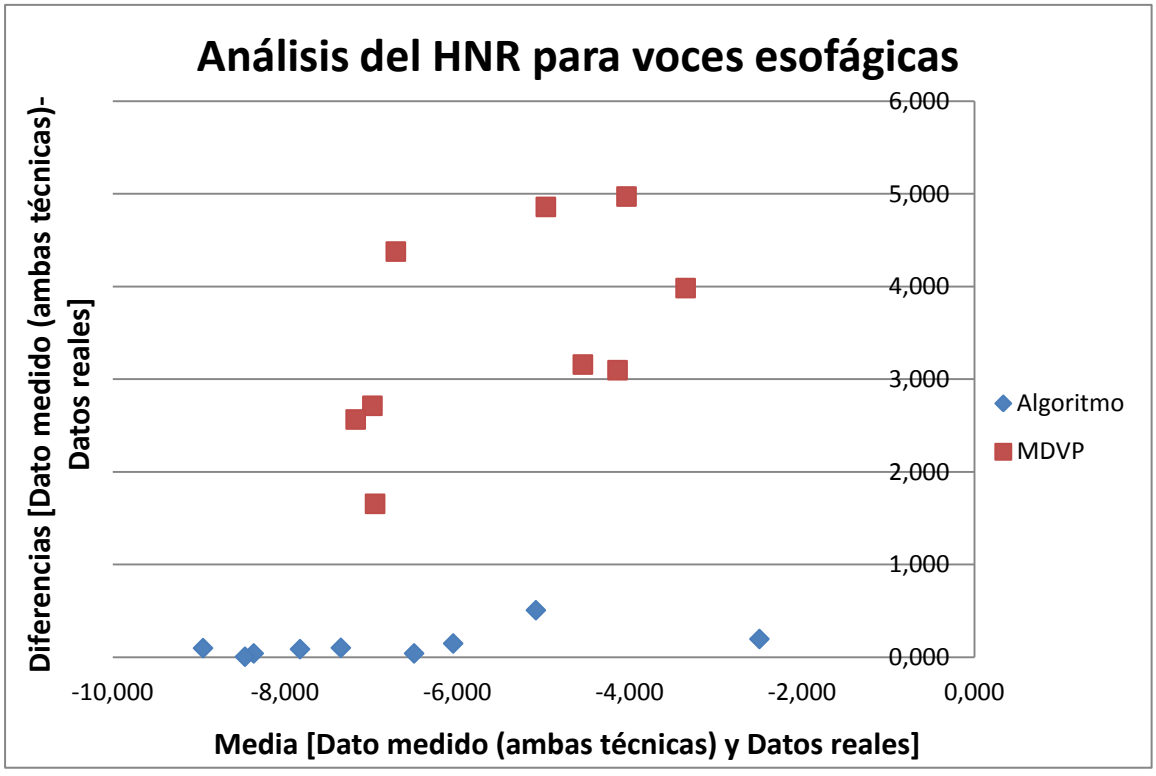

Figura 5.30: Bland-Altman para el HNR de voces esofágicas

[4] ND: No disponible. Para esta voz el programa MDVP no es capaz de mostrar ningún resultado.

En esta ocasión se aprecia claramente que las diferencias entre el MDVP y los datos reales están más alejados del eje de las abscisas que las diferencias entre el algoritmo propuesto y los datos reales. Esto muestra que, en principio, el medidor con el algoritmo propuesto da mejores resultados que el del MDVP para las voces esofágicas, pero esto deberemos probarlo con en el análisis estadístico.

Llevando a cabo un análisis más riguroso, comparamos las medias de las diferencias entre cada uno de los medidores y los datos reales. En esta ocasión, tampoco se puede asumir la normalidad de los datos y, por tanto, se realiza la prueba de Wilcoxon [Wilcoxon45]. Ésta muestra una significancia de p=0,008, lo cual supone que se rechaza la hipótesis nula que dice que las medias de las diferencias entre ambos medidores y las medidas reales son iguales. De hecho, son significativamente diferentes con un 99,2% (>95%). La media de las diferencias entre el algoritmo y las mediadas reales es de 0,136 dB y, la media de las distancias entre el MDVP y las medidas reales es de 3,468 dB. Este resultado junto con la Figura 5.30 evidencia que el algoritmo propuesto es significativamente mejor medidor para la medida del HNR para voces esofágicas que el MDVP. Recordamos que también en esta ocasión no se han tenido en cuenta los datos no válidos que ha dado el MDVP.

Como conclusión en las medidas del HNR, se puede decir que el medidor del algoritmo propuesto es significativamente mejor que el medidor del MDVP.

De hecho, como conclusión general de todos los parámetros, se destaca que el medidor del algoritmo propuesto es significativamente mejor que el MDVP para las voces esofágicas. Para las voces sanas, tanto para el jitter como para el HNR es significativamente mejor, aunque para el pitch y para el shimmer no podamos afirmar dicha mejoría con la significancia suficiente.

CONCLUSIONES

6. CONCLUSIONES

En este capítulo se van a exponer las conclusiones de este trabajo de investigación. Para ello, se verificará que se han cumplido los objetivos marcados en el capítulo de "Introducción" del presente escrito. Los resultados de la investigación se han expuesto y comentado en el capítulo anterior y de ahí se extraerán algunas conclusiones. Posteriormente, se mostrará el impacto científico en términos de publicaciones científicas (revistas, congresos, libros, capítulos de libro etc.) y registros de propiedad intelectual. Finalmente, se analizarán las líneas futuras con las que dar continuidad a este trabajo de investigación.

6.1 CONSECUCIÓN DE LOS OBJETIVOS MARCADOS

De todas las conclusiones que se pueden extraer de este trabajo de investigación, la consecuencia más general, es que ***es posible mejorar la calidad de la voz esofágica y caracterizar dicha voz de forma automática utilizando algoritmos de procesado de señal***. Esta conclusión está basada en los resultados del capítulo 5, en los que se ha probado que los algoritmos de procesado de señal mejoran la calidad de la voz esofágica, concretamente, los parámetros que se han trabajado: el shimmer y HNR (Harmonic to Noise Ratio). En el mismo capítulo, se ha demostrado que se puede caracterizar la voz esofágica dando mejores resultados que los ofrece el paquete de software del mercado Multidimensional Speech Program, el Gold Standard. Estos resultados validan la hipótesis general de la tesis y se cumple con el objetivo general propuesto.

Si analizamos el algoritmo de mejora de la voz esofágica, como ya se ha comentado, está compuesto de tres partes: etapa wavelet, filtro de Kalman y estabilización de polos.

La primera de ellas, la etapa en la que se utiliza la transformada wavelet, modifica el espectro temporal de la voz mediante dicha transformada mejorando especialmente el parámetro shimmer y, por ende, mejorando la calidad de la voz. Se ha demostrado que las medianas de las 30 voces originales y de las procesadas no son iguales con una probabilidad del 99,99%. Además, la media del shimmer de las voces originales es de 1,065 dB y, tras esta primera etapa pasa a ser 0,552 dB, mejorando en 0,513 dB. Valores entorno a 0,5 dB son considerados similares a las voces de las personas sanas según algunos estudios [Gonzalez+02] [Moran+06]. Si unimos el hecho de que las medianas de los dos conjuntos de voces son diferentes y, que se ha producido una mejora en la media, se ha probado que sin ninguna duda que ha habido una mejora en el parámetro shimmer en esta primera etapa logrando uno de los objetivos de esta tesis. Se debe mencionar, con respecto al parámetro HNR, que en esta primera etapa mejora en un 1,157dB.

Los resultados de la segunda etapa están centrados en la mejora del HNR. Se ha demostrado que las medias de las 30 voces anteriores y posteriores a la etapa no son iguales con una probabilidad del 99,99%. Además, la media del HNR de las voces anteriores a la segunda etapa es de -4,996 dB y, tras el filtrado de Kalman pasa a ser -3,548 dB, mejorando en 1,448 dB. Estos dos resultados demuestran que hay una mejora del parámetro HNR en la etapa del filtrado de Kalman. Por lo tanto, se cumple el objetivo de diseñar un algoritmo que disminuya el ruido de la señal de voz esofágica mediante el filtrado de Kalman. En esta etapa el shimmer no sufre cambios significativos.

La tercera etapa, la de estabilización de polos, también disminuye el ruido de la voz esofágica considerablemente. Las medias del parámetro HNR de las voces anteriores y posteriores de la tercera etapa son diferentes en un 99,99%. Si a este

resultado le unimos que la media del HNR de las voces anteriores a la tercera etapa es de -3,548 dB y, tras la etapa de estabilización de polos pasa a ser -2,694 dB, mejorando en 0,764 dB. Esto hace que se disminuya el ruido de las voces esofágicas mejorando así su calidad e inteligibilidad. Al igual que en la etapa anterior, el shimmer no tiene variaciones relevantes. Con todo ello, se consuma el objetivo de concatenar el algoritmo de estabilización de polos y con las otras dos etapas.

Teniendo en cuenta las tres etapas, se produce una mejora media en el HNR de 3,459 dB y de 0,576 dB en el shimmer. Se puede decir, además, que el shimmer ha alcanzado los rangos de normalidad de las voces de las personas sanas. La concatenación de estas tres etapas es la más adecuada. Subjetivamente, el ruido de aspiración del esófago se reduce sustancialmente, como se refleja en un aumento del valor en la prueba de MOS. Por lo tanto, queda demostrado que se ha producido una mejora en la calidad de la voz esofágica.

Analizando ahora el algoritmo de caracterización automática de la voz esofágica, se ha demostrado que el algoritmo propuesto cuantifica objetivamente los parámetros de la voz esofágica, incluso mejor que el Gold standard (Multi-Dimensional Voice Program, MDVP).

Concretamente, se ha demostrado que para los cuatro parámetros estudiados, el pitch, jitter, shimmer y el HNR (Harmonics to Noise Ratio), el algoritmo propuesto es mejor que el MDVP para las voces esofágicas con una significancia mayor del 99%. Es decir, si tenemos en cuenta la distancia a las medidas reales por parte del algoritmo propuesto y por el MDVP, entonces se verifica que el algoritmo propuesto está significativamente más cerca de las medidas reales que las del MDVP. Esto se puede observar en las gráficas de Blant-Altman, mostradas en el apartado 5 de resultados, en las que se percibe claramente que las distancias entre el algoritmo propuesto y las reales son más pequeñas que las distancias entre el MDVP y las medidas reales.

La media del pitch de las distancias entre el algoritmo propuesto y las medidas reales es de 0,735 Hz, mientras que la media con respecto al MDVP es de 10,025 Hz para las voces esofágicas. La media de las distancias con respecto al jitter es de 0,72 % para el algoritmo propuesto y de 2,72 % para el MDVP. En el caso del shimmer, la media para el algoritmo propuesto es de 0,08 dB, mientras que para el MDVP es de 0,63 dB. Finalmente, la media de las distancias del HNR para el algoritmo propuesto es 0,136 dB para el algoritmo propuesto y de 3, 468 dB para el MDVP.

Para las voces sanas, se podría concluir que el algoritmo propuesto es tan eficiente como el Gold estándar, si no mejor en algunos de los casos. Los resultados revelan que para los parámetros del jitter y HNR el algoritmo propuesto es significativamente mejor. Concretamente, las medias de las distancias de las medidas reales con respecto al algoritmo propuesto y al MDVP son significativamente diferentes en un porcentaje superior al 99%. Este hecho, junto con las gráficas de Blant-Altman de los respectivos parámetros, revela sin ninguna duda que el algoritmo propuesto es mejor que el MDVP.

No se puede afirmar, sin embargo, la superioridad del algoritmo propuesto en el caso del pitch y del shimmer. Si bien los datos revelan que para el pitch el algoritmo propuesto parece mejor medidor, no se puede ratificar con una significancia mayor al 95%. En el caso del shimmer la certificación de cuál de los dos medidores es mejor arroja más dudas. Esto no significa, en absoluto, que el algoritmo propuesto no mida bien las voces sanas para el pitch y el shimmer, si no que es tan bueno como el MDVP y, no se puede afirmar que es significativamente mejor. Sabemos, que el Gold standard es un buen medidor para las voces sanas y, se puede decir que el algoritmo propuesto es tan buen medidor como el MDVP.

6.2 IMPACTO CIENTÍFICO

El trabajo de investigación abordado en esta tesis ha dado lugar a publicaciones científicas, registros de propiedad intelectual y algunos proyectos de investigación.

6.2.1 Publicaciones científicas

A continuación, se presentan las revistas científicas en las que se ha publicado los resultados de esta tesis.

Tabla 6.1: Publicación en revistas científicas

Revista	Detalle
Technology and Health Care, 2015	Oleagordia-Ruiz, Ibon; García-Zapirain, Begonya Título: "Harmonic to Noise Ratio Improcement in Oesophageal Speech". (Aprobado, pendiente de publicación) Factor de impacto: 0,636
International Journal of Science and Advanced Technology (IJSAT), 2012	Ruiz, I., García, B., Méndez Título: "Using Games to Assess Oesophageal Voice" [Ruiz+12a]
Computers in Biology and Medicine (CBM), 2009	García, B., Ruiz, I., Méndez, A., Mendezona, M. Título: "Objective Characterization of Oesophageal Voice Supporting Medical Diagnosis, Rehabilitation and Monitoring". Factor de impacto: 1,272. [García+09]
WSEAS Transactions on Systems, 2008	García, B., Ruiz, I., Méndez, A., Mendezona, M. Título: "Oesophageal Voice acoustic Parameterization by means of Optimum Shimmer Calculation". [García+08a]

Además, los resultados de este trabajo de investigación se han publicado parcialmente en los siguientes libros y capítulos de libro.

Tabla 6.2: Libro y capítulos de libro publicados

Libro/Capítulo de libro	Detalle
Wavelet Transforms and their Recent Applications in Biology and Geoscience, 2012	Ruiz, I., García, B. Título: "Improvement of Shimmer Parameter of Oesophageal Voices Using Wavelet Transform" [Ruiz+12b]
Wavelet Theory and their Applications in Engineering, Physics and Technology, 2012	García, B., Ruiz, I. Título: "Oesophageal Speech's Formants Measurement Using Wavelet Transform" [García+12a]
Speech Processing and Auditory Processing Disorders, 2012	García, B., Ruiz, I. Título: "Oesophageal Voice: Objective Quality Assessment" [García+12b]
Recent Advances in Signal Processing 2009	Vicente, J., García, B., Ruiz, I., Méndez, A. Título: "Audio and Image Processing Easy Learning for Engineering students using EasyPAS Tool" [Vicente+09]
La voz esofágica, 2008	García, B., Vicente, J., Ruiz, I., Méndez, A., Mendezona, M. Título: "La voz esofágica. Evaluación objetiva en procesos de diagnóstico, rehabilitación y aprendizaje" [García+08b]

Además, se han presentado varias ponencias a congresos. A continuación se presentas los más recientes:

Tabla 6.3: Publicaciones en congresos más recientes

Congreso	Detalle
BioMed 2014	Oleagordia, I., García, B Título:"Enhancement of Shimmer and HNR in Oesophageal Speech" [Oleagordia+14]
ISSPIT 2013	Ruiz, I., García, B. Título:"Enhancement of Shimmer in Oesophageal Speech Using Different Wavelets" [Ruiz+13]
SIIE 2012	Ruiz, I., García, B. Título: "Enhancement of Shimmer in Oesophageal Speech" [Ruiz+12c]
IWANN 2011	Azzouz, M., García, B., Ruiz, I., Méndez, A. Título:"Oesophageal Voice Harmonic to Noise Ratio Enhancement over UMTS Networks Using Kalman-EM" [Azzouz+11]

ISSPA 2010	Ruiz, I., García, B., Méndez, A. Título: "Two New Approaches of Kalman Filtering for Oesophageal Speech" [Ruiz+10a]
ISIVC 2010	Ruiz, I., García, B., Méndez, A. Título: "New Approach for Oesophageal Speech Enhancement" [Ruiz+10b]
CGAMES 2010	Ruiz, I., García, B., Méndez, A. Título: "Using Games to Assess Oesophageal Voice" [Ruiz+10c]
CISP 2009	Ruiz, I., García, B., Méndez, A. Título: "Two Approaches of Kalman Filtering for Oesophageal Speech" [Ruiz+09]
ICASSP 2008	García, B., Ruiz, I., Méndez, A. Título: "Oesophageal Speech Enhancement Using Poles Stabilization and Kalman Filtering" [García+08c]
EUSIPCO 2008	García, B., Ruiz, I., Méndez, A., Vicente, J. Título: "Extension of EasyPAS Software for the Learning of Image and Audio Digital Processing" [García+08d]
ISIVC 2008	García, B., Ben Jebara, S., Ruiz, I., Mendezona, M. Título: "Oesophageal Voice Quality Assessment Protocol Using Acoustical, Perceptual and Medical Parameters" [García+08e]
ISPRA 2008	Ruiz, I., García, B., Méndez, A., Mendezona, M. Título: "Oesophageal Voice Cycle Detection in Shimmer Calculation Algorithm" [Ruiz+08]
EUSIPCO 2007	García, B., Ruiz, I., Méndez, A., Vicente, J., Mendezona, M. Título:" Automated Characterization of Esophageal and Severely Injured Voices by Means of Acoustic Parameters" [Ruiz+07a]
ISSPIT 2007	Ruiz, I., García, B., Méndez, A., Villanueva, V. Título: "Oesophageal Speech Enhancement Using Kalman Filters" [Ruiz+07b]
BioSignal 2006	Ruiz, I., García, B., Vicente, J.,Méndez, A. Título: "Improvement of the shimmer of esophageal voices using Wavelet" [Ruiz+06]
ISSPIT 2005	García, B., Vicente, J., Ruiz, I., Alonso, A. Título: "Multiplatform Interface Adapted to Pathological Voices" [García+05b]
ICASSP 2005	García, B., Vicente, J., Ruiz, I., Alonso, A., Loyo, E. Título: "Esophageal Voices: Glottal Flow Regeneration" [García+05a]

BIOMED 2005	García, B., Vicente, J., Ruiz, I., Alonso, A., Loyo, E. Título: "Regeneration Model For Esophageal Voices" [García+05c]

También se espera seguir publicando los resultados posteriormente a la entrega de esta tesis.

6.2.2 Propiedad intelectual

A consecuencia de los resultados de esta tesis, se han podido realizar registros de propiedad intelectual, los cuales se presentan a continuación:

- Registro de la base de datos

  Autores: Ibon Oleagordia Ruiz, María Begoña García Zapirain

  Nº de Asiento Registral: 01 / 2013 / 363

  Título del registro de la propiedad intelectual: Grabaciones de Audio con Voz Esofágica del Fonema "a"

  Fecha de presentación y efectos: 14/09/2012

  Clase de obra: Base de datos

- Registro de paquete de software

  Autores: María Begoña García Zapirain, Amaia Méndez Zorrilla, Ibon Ruiz Oleagordia, Mikel Mendezona Goyarzu

  Nº de Asiento Registral: 00 / 2009 / 4935

  Título del registro de la propiedad intelectual: PASVOICE: Software de análisis y procesado de voces esofágicas y laringadas con patologías graves

  Fecha de presentación y efectos: 17/03/2009

  Clase de obra: Programa de ordenador

- Registro de paquete de software

  Autores: María Begoña García Zapirain, Amaia Méndez Zorrilla, Ibon Ruiz Oleagordia, Agustín María Pérez Izquierdo

  Nº de Asiento Registral: 00 / 2009 / 4936

  Título del registro de la propiedad intelectual: ANALISISVOX: Software de ayuda al diagnóstico de patologías en las cuerdas vocales según modelo objetivo.

  Fecha de presentación y efectos: 17/03/2009

  Clase de obra: Programa de ordenador

6.2.3 Proyectos de investigación relacionados con la investigación

Algunos de los resultados del trabajo de investigación de esta tesis han sido obtenidos durante el transcurso de varios proyectos de investigación. En algunos de ellos han colaborado asociaciones y/o empresas que son participantes y, al mismo tiempo, beneficiarias de los resultados. En este sentido, la aportación de dichos agentes ha sido de un gran valor. En concreto, cabe destacar el apoyo obtenido de la "Asociación Vizcaína de Laringectomizados (AVL)" que gracias a sustento se ha podido llevar a cabo esta investigación y se ha podido grabar toda la base de datos de voz esofágica.

Entre los proyectos que han transcurrido en paralelo con esta investigación podemos destacar:

- **Internacionales**

  Aquellos de carácter internacional son: Oesovox (2006-2008) (EUROMED, Institut National de Recherche en Informatique et en Automatique, INRIA); Software para la voz esofágica (2007-2008) Proyectos de Cooperación Interuniversitaria, Túnez-España, Ministerio de Asuntos Exteriores y Cooperación, MAEC); Mejora de las comunicaciones

telefónicas para personas con discapacidad en el habla (Proyectos de Cooperación Interuniversitaria, Túnez-España, Ministerio de Asuntos Exteriores y Cooperación, MAEC); Evaluación objetiva de patologías vocales en base a criterios acústicos y de modelado gráfico (2007-2008) (Proyectos de Cooperación Interuniversitaria, Marruecos-España, Ministerio de Asuntos Exteriores y Cooperación, MAEC).

- **Nacionales**

  Aquellos de ámbito nacional son: Darevoz (2007-2009)- Diagnóstico remoto por la voz a partir de medidas biométricas y otras parametrizaciones (Ministerio de Ciencia e Innovación, MICINN); Dravoes (2009-2010) - Desarrollo del Diagnóstico, Rehabilitación y Aprendizaje de la Voz Esofágica a través de las TICs (Plan Avanza, Ministerio de Industria Trabajo y Comercio, MITyC).

- **Regionales**

  Los que se circunscriben en el ámbito regional son: Esofatic (2008-2009) (Programa Innotek, Departamento de Industria, Comercio y Turismo del Gobierno Vasco (GV-EJ)); Regeneración de la voz esofágica (2006-2007) (Universidad de Deusto); Mediproc (2005-2006) - Evaluación objetiva de la Evolución de las Enfermedades de la Voz (Programa Saiotek, Departamento de Industria, Comercio y Turismo del Gobierno Vasco (GV-EJ)); Larphone (2004-2005) - Mejora de la inteligibilidad en las comunicaciones telefónicas entre laringectomizados (Programa Saiotek, Departamento de Industria, Comercio y Turismo del Gobierno Vasco (GV-EJ)) y Esoimprove (2004-2005) - Sistema de regeneración esofágica (Programa Saiotek, Departamento de Industria, Comercio y Turismo del Gobierno Vasco (GV-EJ)).

6.3 LÍNEAS FUTURAS DE INVESTIGACIÓN

La investigación abierta en esta tesis supone un gran paso en la aplicación de las nuevas tecnologías en la problemática de la baja inteligibilidad de la voz de los laringectomizados y a su parametrización. El trabajo presentado aborda la mejora de la voz en los parámetros Shimmer y HNR, y la parametrización de la voz esofágica. Pero teniendo en cuenta los resultados obtenidos se puede continuar con la investigación ya que varias vías quedan abiertas.

En cuanto a la mejora de la voz, de los cuatro parámetros principales que caracterizan la voz, pitch, jitter, shimmer y HNR, este trabajo ha abordado los dos últimos.

En relación al pitch, se observa que el pitch es mucho menor en las voces esofágicas que en las sanas. Pocos trabajos se han publicado en el sentido de elevación del pitch para voces esofágicas [García03]. Por eso, una línea de investigación sería el diseño de un algoritmo de mejora, en cuanto al citado, de la calidad de las voces esofágicas.

De la misma manera, el parámetro jitter que está relacionado con la variación ciclo a ciclo de la frecuencia fundamental, tiene un valor medido para las voces esofágicas muy por encima que el de las voces sanas o laringadas. Sería deseable que se investigara sobre un algoritmo que modificara, tanto temporal como frecuencialmente, las señales de voz esofágica en este sentido.

Además del análisis propuesto para estos dos nuevos parámetros también se pueden diseñar algoritmos que mejoren más los parámetros abordados en esta tesis. En concreto, el parámetro más susceptible de mejora es el HNR. Se pueden utilizar todo tipo de algoritmos, incluso la combinación de ellos para mejorar este parámetro, por ejemplo, utilizar una codificación de sub-banda, algoritmos neuronales etc.

Otra línea de investigación abierta podría ser el diseño de un algoritmo para la regeneración esofágica que mejore o adapte el modelado de la voz utilizado en este trabajo de investigación, el modelo LPC.

Siguiendo en esta línea, se podría diseñar un algoritmo que mejore el modelo de señal de excitación glótica, o se podría realizar una combinación de la mejora de ambos modelos. Se podría comenzar con la comparativa de la excitación glótica para voces sanas y laringadas realizando el proceso inverso del modelo partiendo de la voz a analizar. De la comparación se deberían extraer unos objetivos a mejorar y, con ellos, se pasaría a diseñar el algoritmo de regeneración de la voz esofágica caracterizada según el modelo de voz elegido. Después vendrían las pruebas y objetivación de resultados, para lo cual convendría que se utilizara el grupo de parámetros de caracterización de la voz más aceptado por la comunidad científica: pitch, jitter, shimmer y HNR.

Por otro lado, como continuación de la línea de investigación afrontada en esta tesis, sería conveniente testear el comportamiento del algoritmo propuesto para las demás vocales. Este algoritmo debe funcionar de forma similar para los fonemas sonoros pero faltaría la comprobación de dicho análisis. En este sentido, posteriormente, se daría el salto a las consonantes sonoras. El objetivo sería estudiar las diferencias que existen entre las voces esofágicas y laringadas y, después, observar el comportamiento de los algoritmos desarrollados para dichas consonantes.

Una vez realizado el estudio de los fonemas sonoros sería adecuado abordar la problemática de los fonemas sordos. Al igual que en el proceso anterior, el primer objetivo sería examinar las diferencias en los parámetros de las voces sanas y esofágicas. Después de detectar qué parámetros son los susceptibles de mejora, habría que diseñar un algoritmo que mejorase la inteligibilidad de los fonemas sordos de la voz esofágica. Es muy posible que los parámetros a mejorar en los fonemas sordos sean diferentes ya que no se contarán con ciclos de la voz y, por tanto, la mejora se centrará en la mejora del ruido.

En cuanto a la parametrización de la voz esofágica, los resultados de esta tesis han sido bastante satisfactorios. El algoritmo propuesto es tan buen medidor o mejor que el Gold Standard. Las nuevas vías de investigación en este sentido se centrarían en la estimación de nuevos parámetros de la voz. Estas medidas podrían ser los relacionados con los cuatro parámetros utilizados en esta investigación y que se han descrito en el capítulo 2 de esta tesis, o bien otros parámetros de la voz.

Por lo tanto, para finalizar, se puede decir que existen varias vías de investigación abiertas tanto para la mejora como para la parametrización de la voz esofágica y, con ello para el desarrollo de proyectos de aplicación de dicha investigación.

En definitiva, se puede decir que son varias las líneas que quedan abiertas, tanto para la investigación sobre mejora de la calidad de la voz esofágica evaluable a través de sus parámetros, como para el desarrollo de proyectos de aplicación de dicha investigación.

CONCLUSIONS

7. CONCLUSIONS

In this chapter, the conclusions drawn from the research work are going to be provided. To this end, the objectives set out in the "Introduction" chapter shall be verified as having been met. The results obtained from the research have been provided and commented on in the previous section. Hence, some conclusions will now be drawn. Subsequently, the scientific impact of this work will be shown, in terms of scientific publications (journals, conferences, books, book chapters, etc.) and intellectual property entries in registries. Lastly, future lines will be analysed, in order to lend continuity to this research work.

7.1 FULFILLMENT OF THE AGREED OBJECTIVES

From all the conclusions that may be drawn from this research work, the most direct consequence is that ***it is possible to improve the quality of the oesophageal voice, and to characterise such a voice automatically, using signal processing algorithms***. This conclusion is based on the results provided in chapter 5, in which it was ascertained that signal processing algorithms improve the quality of the oesophageal voice. In particular, the parameters worked with the Shimmer and HNR (Harmonic to Noise Ratio). In the same chapter, it has been ascertained that the oesophageal voice can be characterised, thus providing better results than the ones offered by the Multidimensional Speech Program market software, the Gold Standard. These results confirm the general hypothesis put forward this thesis, and the proposed general objective has therefore been met.

In analysing the oesophageal voice improvement algorithm, it can be stated that it is composed of three parts: wavelet stage, Kalman filter and pole stabilization.

The first one, the stage in which the wavelet transform is used, modifies the temporary spectrum of the voice by means of this transform, improving the shimmer parameter in particular, and, consequently, improving the equality of the voice. It has been ascertained that the means for the 30 original voices and the processed ones are not the same, with a 99.99% probability. Furthermore, the mean of the shimmer in the original voices is 1.065 dB which, after this first stage, becomes 0.552 dB, showing an improvement of 0.513 dB. Those values close to 0.5 dB are considered similar to the voices of healthy people, according to some studies [Gonzalez+02] [Moran+06]. If we add to this the fact that the mean of the two sets of voices is different, and the fact that there has been an improvement in the mean, it has been ascertained without any doubt that there has been an improvement in the shimmer parameter in this first stage, thus achieving one of the objectives set out of this thesis. As far as the HNR parameter is concerned, it should be mentioned that in this first stage, it improves by 1.157 dB.

The results of the second stage focus on improvement in the HNR. It has been ascertained that the means of the 30 voices prior and subsequent to this stage are not the same, with a 99.99% probability. Additionally, the mean of the HNR of the voices prior to the second stage is -4.996 dB and, after the Kalman filtering process, becomes -3.548 dB, improving by 1.448 dB. These two results show that there is an improvement in the HNR parameter in the Kalman filtering stage. Hence, the objective of designing an algorithm that reduces the noise of the oesophageal voice signal through the Kalman filter is met. In this stage, the shimmer does not undergo significant changes.

The third stage the pole stabilization also significantly reduces the noise in an oesophageal voice. The means of the HNR parameter of the voices previous and subsequent to the third stage differ by 99.99%. If we add to this result the fact that the mean of the HNR of the voices prior to the third stage is

-3.548 dB and, after the pole stabilization stage becomes -2.694 dB, it improves by 0.764 dB. This leads to a reduction in the noise of oesophageal voices, thus improving their quality and intelligibility. Equal to the previous stage, the shimmer does not evidence any relevant variations. However, the aim of concatenating the pole stabilization algorithm with the other two stages is achieved.

Taking into account the three stages, there is an average improvement of 3.459 dB in the HNR and 0.576 dB in the shimmer. It can also be stated that, the shimmer reached normal ranges for healthy people's voices. The concatenation of these three stages is the most appropriate one. Subjectively, the oesophageal breathing noise is reduced substantially, as reflected in the MOS test. It has therefore been demonstrated that there is an improvement in the quality of the oesophageal voice.

Turning now to the automatic characterisation algorithm of the oesophageal voice, it has been ascertained that the proposed algorithm objectively quantifies the oesophageal voice parameters even better than the Gold standard (Multi-Dimensional Voice Program, MDVP).

More specifically, it has been shown that, for the four parameters studied- pitch, jitter, shimmer y el HNR (Harmonics to Noise Ratio)- the proposed algorithm is better than the MDVP for oesophageal voices, with a significance greater than 99%. In other words, if we take into account the distance to the real measurements by the proposed algorithm and by the MDVP, it can be then ascertained that the proposed algorithm is significantly closer to the real measurements than the MDVP. This can be observed in Blant-Altman graphics, shown in section 5 of the results chapter, in which it can be clearly perceived that the distances between the proposed algorithm and the real ones are smaller than the distances between the MDVP and the real measurements.

The average pitch of the distances between the proposed algorithm and real measurements is 0.735 Hz, whereas the mean in terms of the MDVP is

10.025 Hz for oesophageal voices. The mean for the distances with regard to jitter is 0.72 % for the proposed algorithm, and 2.72 % for MDVP. In the case of the shimmer, the mean for the proposed algorithm is 0.08 dB, whereas for the MDVP is 0.63 dB. Lastly, the mean for the distances of HNR for the proposed algorithm is 0.136 dB and 3.468 dB for the MDVP.

For healthy voices, it can be concluded that the proposed algorithm is as efficient as the Gold Standard, if not better in some cases. The results reveal that for jitter and HNR parameters, the proposed algorithm is significantly better. In particular, the means for the distances of the real measurements with regard to the proposed algorithm and the MDVP are differ significantly by a percentage higher than 99%. This fact, together with Blant-Altman graphics of the respective parameters, undoubtedly shows that the proposed algorithm is better than the MDVP.

The superiority of the proposed algorithm cannot, therefore, be confirmed in the case of the pitch and shimmer. While data reveals that for the pitch the proposed algorithm seems to be the best gauge, it cannot be ratified with significance higher than 95%. In the case of the shimmer, the proof as to which of the two gauges is the best one casts further doubts. This on no account means that the proposed algorithm does not correctly measure the healthy voices for the pitch and shimmer, although it means that it is as good as the MDVP, and it cannot be stated that it is significantly better. We do know that Gold standard is a good gauge for healthy voices, and it can be said that the proposed algorithm is as a good gauge as the MDVP.

7.2 SCIENTIFIC IMPACT

The research discussed in this thesis has resulted in scientific publications, intellectual property entries in registries and certain research projects.

7.2.1 Scientific publications

The scientific publications in which the results of this thesis have been published are as follows:

Table 7.1: Scientific journals

Journal	Detail
Technology and Health Care, 2015	Oleagordia-Ruiz, Ibon; García-Zapirain, Begonya Título: "Harmonic to Noise Ratio Improcement in Oesophageal Speech". (Approved) Impact factor: 0.636
International Journal of Science and Advanced Technology (IJSAT), 2012	Ruiz, I., García, B., Méndez Title: "Using Games to Assess Oesophageal Voice" [Ruiz+12a]
Computers in Biology and Medicine (CBM), 2009	García, B., Ruiz, I., Méndez, A., Mendezona, M. Title: "Objective Characterization of Oesophageal Voice Supporting Medical Diagnosis, Rehabilitation and Monitoring". Impact factor: 1.272. [García+09]
WSEAS Transactions on Systems, 2008	García, B., Ruiz, I., Méndez, A., Mendezona, M. Title: "Oesophageal Voice acoustic Parameterization by means of Optimum Shimmer Calculation". [García+08a]

Furthermore, the results of this research have been partially published in the following books and book chapters.

Table 7.2: Books and book chapters

Book/Book Chapter	Detail
Wavelet Transforms and their Recent Applications in Biology and Geoscience, 2012	Ruiz, I., García, B. Title: "Improvement of Shimmer Parameter of Oesophageal Voices Using Wavelet Transform" [Ruiz+12b]
Wavelet Theory and their Applications in Engineering, Physics and Technology, 2012	García, B., Ruiz, I. Title: "Oesophageal Speech's Formants Measurement Using Wavelet Transform" [García+12a]
Speech Processing and Auditory Processing Disorders, 2012	García, B., Ruiz, I. Title: "Oesophageal Voice: Objective Quality Assessment" [García+12b]
Recent Advances in Signal Processing 2009	Vicente, J., García, B., Ruiz, I., Méndez, A. Title: "Audio and Image Processing Easy Learning for Engineering students using EasyPAS Tool" [Vicente+09]
La voz esofágica, 2008	García, B., Vicente, J., Ruiz, I., Méndez, A., Mendezona, M. Title: "La voz esofágica. Evaluación objetiva en procesos de diagnóstico, rehabilitación y aprendizaje" [García+08b]

In addition, several international conference papers have been presented.

Table 7.3: International conference

Conference	Detail
BioMed 2014	Oleagordia, I., García, B Title:"Enhancement of Shimmer and HNR in Oesophageal Speech" [Oleagordia+14]
ISSPIT 2013	Ruiz, I., García, B. Title:"Enhancement of Shimmer in Oesophageal Speech Using Different Wavelets" [Ruiz+13]

SIIE 2012	Ruiz, I., García, B. Title: "Enhancement of Shimmer in Oesophageal Speech" [Ruiz+12c]
IWANN 2011	Azzouz, M., García, B., Ruiz, I., Méndez, A. Title:"Oesophageal Voice Harmonic to Noise Ratio Enhancement over UMTS Networks Using Kalman-EM" [Azzouz+11]
ISSPA 2010	Ruiz, I., García, B., Méndez, A. Title: "Two New Approaches of Kalman Filtering for Oesophageal Speech" [Ruiz+10a]
ISIVC 2010	Ruiz, I., García, B., Méndez, A. Title: "New Approach for Oesophageal Speech Enhancement" [Ruiz+10b]
CGAMES 2010	Ruiz, I., García, B., Méndez, A. Title: "Using Games to Assess Oesophageal Voice" [Ruiz+10c]
CISP 2009	Ruiz, I., García, B., Méndez, A. Title: "Two Approaches of Kalman Filtering for Oesophageal Speech" [Ruiz+09]
ICASSP 2008	García, B., Ruiz, I., Méndez, A. Title: "Oesophageal Speech Enhancement Using Poles Stabilization and Kalman Filtering" [García+08c]
EUSIPCO 2008	García, B., Ruiz, I., Méndez, A., Vicente, J. Title: "Extension of EasyPAS Software for the Learning of Image and Audio Digital Processing" [García+08d]
ISIVC 2008	García, B., Ben Jebara, S., Ruiz, I., Mendezona, M. Title: "Oesophageal Voice Quality Assessment Protocol Using Acoustical, Perceptual and Medical Parameters" [García+08e]
ISPRA 2008	Ruiz, I., García, B., Méndez, A., Mendezona, M. Title: "Oesophageal Voice Cycle Detection in Shimmer Calculation Algorithm" [Ruiz+08]
EUSIPCO 2007	García, B., Ruiz, I., Méndez, A., Vicente, J., Mendezona, M. Title: "Automated Characterization of Esophageal and Severely Injured Voices by Means of Acoustic Parameters" [Ruiz+07a]

ISSPIT 2007	Ruiz, I., García, B., Méndez, A., Villanueva, V. Title: "Oesophageal Speech Enhancement Using Kalman Filters" [Ruiz+07b]
BioSignal 2006	Ruiz, I., García, B., Vicente, J.,Méndez, A. Title: "Improvement of the shimmer of esophageal voices using Wavelet" [Ruiz+06]
ISSPIT 2005	García, B., Vicente, J., Ruiz, I., Alonso, A. Title: "Multiplatform Interface Adapted to Pathological Voices" [García+05b]
ICASSP 2005	García, B., Vicente, J., Ruiz, I., Alonso, A., Loyo, E. Title: "Esophageal Voices: Glottal Flow Regeneration" [García+05a]
BIOMED 2005	García, B., Vicente, J., Ruiz, I., Alonso, A., Loyo, E. Title: "Regeneration Model For Esophageal Voices" [García+05c]

It is also expected to continue publishing the results subsequent to delivery of this thesis.

7.2.2 Intellectual property

Several intellectual property entries in registries have been obtained as a result of this thesis.

- Database entries

Authors: Ibon Oleagordia Ruiz, María Begoña García Zapirain

Entry number: 01 / 2013 / 363

Intellectual property entries title: Grabaciones de Audio con Voz Esofágica del Fonema "a"

Filing Date and Effects: 2012/09/14

Type: Database

- Software entries

Authors: María Begoña García Zapirain, Amaia Méndez Zorrilla, Ibon Ruiz Oleagordia, Mikel Mendezona Goyarzu

Entry number: 00 / 2009 / 4935

Intellectual property entry title: PASVOICE: Software de análisis y procesado de voces esofágicas y laringadas con patologías graves

Filing Date and Effects: 2009/03/17

Type: Software

- Software entries

Authors: María Begoña García Zapirain, Amaia Méndez Zorrilla, Ibon Ruiz Oleagordia, Agustín María Pérez Izquierdo

Entriy number: 00 / 2009 / 4936

Intellectual property entry title: ANALISISVOX: Software de ayuda al diagnóstico de patologías en las cuerdas vocales según modelo objetivo.

Filing Date and Effects: 2009/03/17

Type: Software

7.2.3 Research projects related to the thesis

Some of the research results of this thesis have been obtained over the course of several research projects. Associations and/or companies have collaborated in some of the projects, participating and at the same time benefiting from the results. In this sense, the contribution of these agents has been of great value. In particular, attention should be drawn to the support obtained from the "Asociación Vizcaína de Laringectomizados (AVL)". This research has been

carried out thanks to them and it has been able to record the entire oesophageal voice database.

Among the projects related to this research can be included the following:

- **International**

International projects: Oesovox (2006-2008) (EUROMED, Institut National de Recherche en Informatique et en Automatique, INRIA); "Software para la voz esofágica" (2007-2008); Interuniversity Cooperation Projects, Tunisia-Spain, Ministry of Foreign Affairs and Cooperation); "Mejora de las comunicaciones telefónicas para personas con discapacidad en el habla" (Interuniversity Cooperation Projects, Tunisia-Spain, Ministry of Foreign Affairs and Cooperation); "Evaluación objetiva de patologías vocales en base a criterios acústicos y de modelado gráfico" (2007-2008) (Interuniversity Cooperation Projects, Morocco-Spain, Ministry of Foreign Affairs and Cooperation).

- **National**

National projects: Darevoz (2007-2009) "Diagnóstico remoto por la voz a partir de medidas biométricas y otras parametrizaciones" (Ministry of Science and Innovation); Dravoes (2009-2010) "Desarrollo del Diagnóstico, Rehabilitación y Aprendizaje de la Voz Esofágica a través de las TICs" (Avanza programme, Ministry of Industry, Labor and Trade).

- **Regional**

Regional projects: Esofatic (2008-2009) (Innotek programme, Department of Industry, Trade and Tourism of the Basque Government); "Regeneración de la voz esofágica" (2006-2007) (University of Deusto); Mediproc (2005-2006) - "Evaluación objetiva de la Evolución de las Enfermedades de la Voz" (Saiotek programme, Department of Industry, Trade and Tourism of the Basque Government); Larphone (2004-2005) "Mejora de la inteligibilidad en las comunicaciones telefónicas entre laringectomizados" (Saiotek programme,

Department of Industry, Trade and Tourism of the Basque Government) and Esoimprove (2004-2005) "Sistema de regeneración esofágica" (Saiotek programme, Department of Industry, Trade and Tourism of the Basque Government).

7.3 FUTURE RESEARCH

The research pursued in this thesis constitutes a significant step in implementing new technologies regarding the problem of low intelligibility of oesophageal speech and its parameterization. The work presented here deals with improving the voice in case of the Shimmer and HNR parameters and the parameterization of the oesophageal voice. However, considering the results obtained, the research may be pursued in several ways.

As regards speech enhancement, out of four main parameters referring to the characterisation of speech, pitch, jitter, shimmer and HNR, this work has addressed the latter two.

Regarding pitch, it can be noticed that the pitch is lower in oesophageal speech than in healthy speech. Few works have been published about the elevation of oesophageal voice pitch [García03]. Therefore, one of the future lines of research could be to design a speech enhancement algorithm in order to improve this parameter.

Similarly, the jitter parameter, related to the variation of cycles of fundamental frequency, has a much high mean value for oesophageal speech than healthy speech. It would be desirable to conduct research into an algorithm to modify both the time and frequency domain in terms of oesophageal voice signals in this respect.

Besides these two new parameters, new algorithms could also be designed in order to vastly improve the parameters discussed in this thesis. Specifically, the HNR parameter is capable of improvement. All kinds of algorithms can be used,

even combining them to improve this parameter for example, using subband coding and neural algorithms etc.

Another line of research that can be pursued could be the design of an algorithm that improves the LPC model.

To this end, an algorithm could be designed that improves the glottal excitation signal model, or it could perform a combination of enhancement of both models. The glottal excitation could be compared for laryngeal and healthy speech. Goals to be met should be achieved as a result of such a comparison and, alongside this an algorithm could be design to regenerate the esophageal voice characterised by the chosen model voice. Trials and evaluations would then be carried out. It would be appropriate to use the group of parameters accepted by the scientific community for characterization purpose: pitch, jitter, shimmer and HNR.

As a continuation of the research covered in this thesis, it would be desirable to test the behaviour of the proposed algorithm for other vowels. This algorithm should work similarly for voiced phonemes but lacks the verification of such an analysis. The voiced consonants could then be dealt with, whereby the aim would be to study the differences between oesophageal and laryngeal voices and then observe the behaviour of the algorithms developed for these consonants.

Once the voiced phonemes have been studied, it would then be appropriate to address the unvoiced phoneme problem. As in the above process, the first objective would be to examine the differences in the parameters for healthy and oesophageal speech. Once improvable parameters has been detected, an algorithm would then be designed that would enhance the intelligibility of the unvoiced oesophageal speech phonemes. The parameters to be enhanced for oesophageal unvoiced phonemes would probably be different from voiced phonemes since there are no pitch peaks. Therefore, speech improvement would focus on enhancing the noise.

As for parameterization of the oesophageal voice, the results obtained from this thesis have proved to be quite satisfactory. The proposed algorithm is as good gauge or better than the Gold Standard. The new lines of research in this sense would focus on estimating new voice parameters. These could be related to the four parameters used in this research that have been described in Chapter 2 of this thesis, or other speech parameters.

Therefore, to conclude, we can say that there are several lines of research that could focus both on an improvement in and parameterization of oesophageal voice and thereby application projects could be develop via such research.

In short, we can say that there are several lines open to both research into improving the quality of the evaluable oesophageal voice via parameters and into the development of application projects.

REFERENCIAS BIBLIOGRÁFICAS

8. REFERENCIAS BIBLIOGRÁFICAS

[Abramovich+07] Abramovich Y.I., S. N. (2007). Order Estimation and Discrimination Between Stationary and Time-Varying (TVAR) Autoregressive Models. *IEEE Transactions on Signal Processing , volume 55* (issue 6, part 2), pp. 2861-2876.

[Akansu92] Akansu, A. N. (1992). *Multiresolution signal decomposition: transforms, subbands, and wavelets.* Boston: Academic Press.

[Akansu+10] Akansu, A., Serdijn, W., & Selesnick, I. (2010). Wavelet Transforms in Signal Processing: A Review of Emerging Applications. *Physical Communication , 3* (1), 1-18.

[Allingham+98] Allingham, D., West, M., & Alistair, I. M. (1998). Wavelet Reconstruction of Nonlinear Dynamics. *International Journal of Bifurcation and Chaos , 8* (11), 2191-2201.

[Altman+83] Altman, D., & Bland, J. (1983). Measurement in medicine: the analysis of method comparison studies. *The Statistician* (32), 307-317.

[Azzouz+11] Azzouz, M., García, B., Ruiz, I., & Méndez, A. (2011). Oesophageal Voice Harmonic to Noise Ratio Enhancement over UMTS Networks Using Kalman-EM. *IWANN*, (págs. 265-272). Torremolinos, Spain.

[Bäckström04] Bäckström, T. (2004). *Linear Predictive modelling of Speech-Constraints and Line Spectrum Pair Decomposition. Tesis Doctoral.* Helsinki: Helsinki University of Technology, Espoo, Finland.

[Baken+00] Baken, P. J., & Orlikoff, R. F. (2000). *Clinical Measurement of Speech and Voice.* San Diego: Singular Publishing Group.

[Baker+09] Baker, J. M., Deng, L., Glass, J., Khudanpur, S., Lee, C.-H., Morgan, N., y otros. (2009). Research Developments and Directions in Speech Recognition and Understanding. *IEEE SIGNAL PROCESSING MAGAZINE* , 75-80.

[Barroto+02] Barroto Cruz, R., & Aneiros Ribas, R. (2002). *Investigación - Acción.* La Habana: Escuela Nacional de Investigación Pública.

[Bay99] Bay, J. S. (1999). *Fundamentals of Linear State Space Systems.* McGraw-Hill.

[Boersma93] Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *IFA proceedings 17*, (págs. 97-110).

[Boone97] Boone, D. (1997). The singing/acting voice in the mature adult. *Journal of Voice , 2* (11), 161-164.

[Bourlard+98] Bourlard, H., Kamp, Y., Ney, H., & Wellekens, C. (1988). Speaker dependent connected speech recognition via dynamic programming and statistical methods. *Speech and Speaker Recognition , 12*.

[Brown+92] Brown, R., & Hwang, P. (1992). *Introduction to Random Signals and Applied Kalman Filtering*. John Wiley & Sons.

[Burns07] Burns, D. (2007). Systemic Action Research: A strategy for whole system change. En D. Burns. Bristol: Policy Press.

[Burrus+98] Burrus, C. S., Goinath, R. A., & Guo, H. (1998). INTRODUCTION TO WAVELETS AND WAVELET TRANSFORMS, A PRIMER. Upper Saddle River NJ (USA): Prentice Hall.

[Calderbank+96] Calderbank, A., Daubechies, I., Swedens, W., & Yeo, B. (1996). Wavelet Transforms tha Map Integers to Integers. *Applied and Computational Harmonic Analysis , 5* (3), 332-369.

[Carmi+10] Carmi, A., Gurfil, P., & Kanevsky, D. (2010). Methods for sparse signal recovery using Kalman filtering with embedded pseudo-measurement norms and quasi-norms. *IEEE Transactions on Signal Processing , 58* (4), 2405–2409.

[Chen+01] Chen, J., & Kao, Y. (2001). Pitch marking based on an adaptable filter and a peakvalley estimation method. *Computational Linguistics and Chinese Language Processing , 6*, 1-112.

[Cheveigné+02] Cheveigné, A., & Kawahara, H. (2002). Yin, a fundamental frequency estimator for speech and music, 111 (4) (2002). *Journal of the Acoustical Society of America , 11* (4).

[Chui92] Chui, C. K. (1992). *An Introduction to Wavelets.* San Diego: Academic Press.

[Cohen+92] Cohen, A., Daubechies, I., & Feauveau, J. (1992). Bi-orthogonal bases of compactly supported wavelets. *Commun. Pure Appl. Math. , 45*, 485–560.

[Daubechies88] Daubechies, I. (1988). Orthogonal bases of compactly supported wavelets. *Commun. Pure Appl. Math. , 41*, 909–996.

[Daubechies92] Daubechies, I. (1992). *TEN LECTURES ON WAVELETS.* Philadelphia: 2nd ed. Philadelphia: SIAM.

[DeBonis+08] DeBonis, D. A., & Moncrieff, D. (2008). Auditory Processing Disorders: an Update for Speech-Language Pathologists. *American Journal of Speech-Language Pathology , 17*, 4-18.

[Deliyeski93] Deliyeski, D. D. (1993). MDVP Acoustic Model and Evaluation of Pathological Voice Production. *Eurospeech* .

[Dorken+94] Dorken, E., Nawab, S. H. (1994). Improved musical pitch tracking using principal decomposition analysis. *ICASSP*, (págs. 217-220).

[Doval+91] Doval, B., Rodet, X. (1991). Estimation of fundamental frequency of musical sound signals. *ICASSP*, (págs. 3657–3660).

[Doval+93] Doval, B., & Rodet, X. (1993). Fundamental frequency estimation and tracking using maximum likelihood harmonic matching and HMMs. *ICASSP*, (págs. 221–224).

[Dromey+08] Dromey, C., & Smith, M. E. (2008). Vocal Tremor and Vibrato in the Same Person: Acoustic and Electromyographic Differences. *Journal of Voice , 22* (5), 541 - 545.

[Einicke09] Einicke, G. (2009). Asymptotic Optimality of the Minimum-Variance Fixed-Interval Smoother. *IEEE Trans. Automatic Control, 54* (12), 2904–2908.

[Elemetrics94] Elemetrics, K. (1994). Disordered Voice Database. Massachusetts, Estados Unidos.

[Fant+85] Fant, G., Liljencrants, J., & Lin, Q. (1985). A four-parameter model of glottal flow. *STL-QPRS* , 1-13.

[Fasano+87] Fasano, G., & Franceschini, A. (1987). Two-dimensional goodness-of-fit testing in astronomy. *Monthly Notices of the Royal Astronomical Society* , 155-170.

[Feijoo+90] Feijoo, S., Hernández, C. (1990). Short-term stability measures for the evaluation. *J. Speech. Hear. Res. , 33*, 324-334.

[Feneis94] Feneis, H. (1994). *Nomenclatura anatómica ilustrada.* Barcelona: Salvat editores.

[Ferrand02] Ferrand, C. T. (2002). Harmonics-to-Noise Ratio: An Index of Vocal Aging. *Journal of Voice , 16* (4), 480–487.

[Ferrer+05] Ferrer, C. A., Gonzalez, E., & Hernández-Díaz, M. E. (2005). Correcting the Use of Ensemble averages in the Calculation of Harmonics to Noise Ratios in voice signals. *Journal Acoust. Soc. America* , 605-607.

[Ferrer+06] Ferrer, C. A., E. G.-D. (2006). Evaluation of Time and Frequency Domain-Based Methods for the Estimation of Harmonics-to-Noise-Ratios in Voice Signals. *Progress in Pattern Recognition, Image Analysis and Applications (Lecture Notes in Computer Science) , volume 4225/2006*, pp. 406-415.

[Flanagan65] Flanagan, J. L. (1965). Speech Analysis, Synthesis and Perception. *Springer*.

[Flipsen06] Flipsen, P. (2006). Measuring the intelligibility of conversational speech in children. *Clinical Linguistics & Phonetics , 20* (4), 303-312.

[Fourier22] Fourier, J. (1822). *Théorie Analytique de la Chaleur*. Paris: Chez Firmin Didot, père et fils.

[Friedman39] Friedman, M. (1939). The use of ranks to avoid the assumption of normality implicit in the analysis of variance". Journal of the American Statistical Association. *Journal of the American Statistical Association* , 675–701.

[Gabrea04] Gabrea, M. (2004). ICASSP. *Robust adaptive kalman filtering-based speech enhancement algorithm.*

[Gail11] Gail, R. J. (2011). The Role of the Speech-Language Pathologist in Identifying and Treating Children with Auditory Processing Disorder. *Language, Speech, and Hearing Services in Schools , 42*, 241-245.

[Gannot98] Gannot, S., Burshtein, D., & Weinstein, E. (1998). Iterative and Sequential Kalman Filter Based Speech Enhancement Algorithms. *IEEE Transaction on Speech, Audio Processing , 6* (4), 373 - 385.

[García03] García Zapirain, B. (2003). Modelo de Procesado Digital para la Regeneración de la Voz Esofágica.

[García+05a] García, B., Vicente, J., Ruiz, I., Alonso, A., & Loyo, E. (2005). Esophageal Voices: Glottal Flow Restoration. *ICASSP*, (págs. 141-144).

[García+05b] García, B., Vicente, J., Ruiz, I., & Alonso, A. (2005). Multiplatform Interface Adapted to Pathological Voices. *ISSPIT*, (págs. 912-917). Atenas.

[García+05c] García, B., Vicente, J., Ruiz, I., Alonso, A., & Loyo, E. (2005). Regeneration Model For Esophageal Voices. *Biomed*, (págs. 439-444). Innsbruck.

[García+08a] García, B., Ruiz, I., Méndez, A., & Mendezona, M. (2008). Oesophageal voice acoustic parameterization by means of optimum shimmer calculation. *WSEAS Transactions on Systems , 7*, 489-499.

[García+08b] García, B., Vicente, J., Ruiz, I., Méndez, A., Pérez, A., & Mendezona, A. (2008). *La voz esofágica.* Bilbao: Publicaciones Universidad de Deusto.

[García+08c] García, B., Ruiz, I., & Méndez, A. (2008). Oesophageal Speech Enhancement Using Poles Stabilization and Kalman Filtering. *ICASSP*, (págs. 1597 - 1600). Las Vegas, EE.UU.

[García+08d] García, B., Ruiz, I., Méndez, A., & Vicente, J. (2008). Extension of EasyPAS Software for the Learning of Image and Audio Digital Processing. *European Signal Processing Conference (EUSIPCO).* Lausanne (Suiza).

[García+08e] García, B., Ben Jebara, S., Ruiz, I., & Mendezona, M. (2008). Oesophageal Voice Quality Assessment Protocol Using Acoustical, Perceptual and Medical Parameters. *The International Symposium on Image/Video Communications over fixed and mobile networks. ISIVC 2008* (págs. 7-12). Bilbao: University of Deusto.

[García+09] García, B., Ruiz, I., Méndez, A., & Mendezona, M. (2009). Objective characterization of oesophageal voice supporting medical diagnosis, rehabilitation and monitoring. *Computers in Biology and Medicine , 39/2*, 97-105.

[García+12a] García, B., & Ruiz, I. (2012). Oesophageal Speech's Formants Measurement Using Wavelet Transform. En D. Baleanu, *Advances in Wavelet Theory and their Apllications in Engineering, Phisics and Technology* (págs. 81-98). Croacia: Intech.

[García+12b] García, B., Ruiz, I., Méndez, A., & Mendezona, M. (2012). Oesophageal Voice: Objective Quality Assessment. En P. L. Taryn, & R. E. Lillian, *Speech Processing and Auditory Processing Disorders* (págs. 109-128). New York: Nova Biomedical.

[Gerratt05] Gerratt, J. K. (2005). Perception of aperiodicity in pathological voice. *Journal of the Acoustic Society of America , Volume 117* (Issue 4), pp. 2201-2211.

[Gibiat88] Gibiat, V. (1988). Phase space representations of acoustical musical signals, Journal of Sound and Vibration. *Journal of Sound and Vibration , 123* (3), 537–572.

[Gibson+91] Gibson, J., Koo, B., & Gray, S. (1991). Filtering of colored noise for speech enhancement and coding. *IEEE Trans. Signal Processing , 39*, 1732–1742.

[Giddens97] Giddens, A. (1997). Las nuevas reglas del método sociológico.

[Girin06] Girin, L. (2006). Theoretical and experimental bases of a new method for accurate separation of harmonic and noise components of speech signals. Florence, Italy: European Signal Processing conference (EUSIPCO).

[Goh+99] Goh, K., Tan, C., & Tan, B. (1999). Kalman Filtering Speech Enhancement Method Based on a Voiced-Unvoiced Speech Model. *IEEE Transaction on Speech, Audio Processing , 7* (5), 510 - 524.

[Gonzalez+02] Gónzalez, J., Cervera, T., & Miralles, J. (2002). Análisis acústico de la voz: fiabilidad de un conjunto de parámetros multidimensionales. *Acta Otorrinolaringológica de Esp.* , 256-268.

[Grancharov+06] Grancharov, V., Samuelsson, J., & Kleijn, B. (2006). On Causal Algorithms for Speech Enhancement. *IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING , 14* (3), 764-773.

[Grant+00] Grant, K. W., & Seitz, P. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal Acoustic Sociaty of America* (108), 1197-1208.

[Hagmüller+06] Hagmüller, M., & Kubina, G. (2006). Poincaré pitch marks, 48 (12). *Speech Communication , 48* (12), 1650-1665.

[Harries+98] Harries, M., Hawkins, S., Hacking, J., & Hughes, I. (1998). Changes in the male voice at puberty:vocal folds length and its relationship to the fundamental frequency. *J. Laryngolo. Otol.* , 451-454.

[Jacobs93] Jacobs, O. (1993). *Introduction to Control Theory.* Oxford University Press.

[Jacobsen+03] Jacobsen, E., & Lyons, R. (2003). The sliding DFT. *Signal Processing Magazine , 20* (2), 74-80.

[Javkin+97] Javkin, H., Galler, M., & Niedzielski, N. (1997). Enhancement of esophageal speech by injection noise rejection. *IEEE Proc.*, (págs. 1207-1210).

[Kadambe+91] Kadambe, S., & Bourdreaux-Bartels, G. F. (1991). A Comparison of a Wavelet Functions for Pitch Detection of Speech Signals. *ICASSP* , (págs. 449-452).

[Kadambe+92] Kadambe, S., & Bourdreaux-Bartels, G. F. (1992). Application Of The Wavelet Transform For Pitch Detection Of Speech Signals. *IEEE Transaction On Information Theory* (38), 917-924.

[Kalman60] Kalman, R. (1960). A New Approach to Linear Filtering and Prediction Problems. *Transaction of the ASME - Journal of Basic Engineering* , 35-45.

[Karthikeyan+05] Karthikeyan Umapathy, S. K. (2005). Discrimination of pathological voices using a time-frequency approach. *IEEE Transactions on Biomedical Engineering , vol. 52* (no. 3), pp. 421-430.

[Kasuya+86] Kasuya, H., Ogawa, S., Mazuma, K., Ebihara, S. (1986). Normalized noise energy as an acoustic measure to evaluate pathologic voice. *J. Acoust. Soc. Am. , 80* (5), 1329-1334.

[Kearney04] Kearney, A. (2004). Esophageal Speech. *Otolaryngol Clin N Am* , 613-625.

[Kedem86] Kedem, B. (1986). Spectral analysis and discrimination by zero-crossings. *Proceedings of the IEEE , 74* (11), 1477-1493.

[Kim+10] Kim, Y.-J., Beutnagel, C. (2010). Automatic detection of abnormal stress patterns in unit selection synthesis. *Proceedings Interspeech.*

[Kojima+80] Kojima, H., Gould, W. J., Lambiase, A., & Isshiki, N. (1980). Computer Analysis of Hoarseness. *Acta Otolaringol.* (89), 547-554.

[Krishnamurthy92] Krishnamurthy, A. (1992). Glottal source estimation using a sum of exponentials model. *IEEE Transactions on Signal Processing* , 682-686.

[Krom94] Krom, G. d. (1994). *Acoustics Correlates of Breathiness and Roughness (PhD-Thesis)*. Utrecht: LEd.

[Kumara07] Kumara Shama, A. K. (2007). Study of harmonics-to-noise ratio and critical-band energy spectrum of speech as acoustic indicators of laryngeal and voice pathology. *EURASIP Journal on Applied Signal Processing , volume 2007* (issue 1), pp. 50-59.

[Laitman86] Laitman, J. (1986). El origen del lenguaje articulado. *Mundo científico , 6* (64), 1182-1191.

[Levinson47] Levinson, N. (1947). The Wiener (RMS) error criterion in filter design and prediction. *Journal of Mathematics and Physics , 25* (4), 261-278.

[Lewin46] Lewin, K. (1946). Action Research and Minority Problems. *Journal of Social Issues* , 34-36.

[Lewis+08] Lewis, F., Xie, L., & Popa, D. (2008). *Optimal and Robust Estimation.* Broken Sound Parkway NW: CRC Press.

[Lim+78] Lim, J., & Oppenheim, A. (1978). All-pole modeling of degraded speech. *IEEE Trans. Acoustic, Speech, Signal Processing* , 191–210.

[Lió03] Lió, P. (2003). Wavelets in bioinformatics and computational biology: state of art and perspectives. *BIOINFORMATICS REVIEW , 19* (1), 2-9.

[Love+08] Love, D., Heath, R., Lau, V., Gesbert, D., Rao, B., & Andrews, M. (2008). An overview of limited feedback in wireless communication systems. *IEEE Journal on Selected Areas Communications , 26*, 1341–1365.

[Mallat89] Mallat, S. (1989). A Theory for Multiresolusion Signal Decomposition: The Wavelet Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence , 11* (7), 674- 693.

[Mallat99] Mallat, S. (1999). *A Wavelet Tour of Signal Processing.* A. Press.

[Michaelis+98] Michaelis, D., Fröhlich, M., & Strube, H. (1998). Selection and combination of acoustic features for the description of pathologic voices. *J. Acoust. Soc. Am. , 103* (3), 1628-1639.

[Michaelis+97] Michaelis, D., Gramms, T., & Strube, H. (1997). Glottal-to-Noise Excitation Ratio – a New Measure for Describing Pathological Voices. *Acta Acustica , 83*, 700-706.

[Moran+06] Moran, R. J., Reilly, R. B., Chazal, P. d., & Lacy, P. D. (2006). Telephony-Based Voice Pathology Assessment Using. *IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING , 53* (3), 468-477.

[Murphy06] Murphy, P. J. (2006). On First Rahmonic Amplitude in the Analysis of Synthesized Aperiodic Voice Signals. *Journal Acoust. Soc. Am. , 120* (5), 2896-2907.

[Murphy+07a] Murphy, P.J., O. O. (2007). Cepstrum-Based Estimation of the Harmonics-to-Noise Ratio for Synthesized and Human Voice Signals. En *Nonlinear Analyses and Algorithms for Speech Processing (Lecture Notes in Computer Science. Volume 3817/2005)* (págs. pp. 150-160). Springer Berlin / Heidelberg.

[Murphy+07b] Murphy, P.J., O. O. (2007). Noise estimation in voice signals using short-term cepstral analysis. *Journal of the Acoustical Society of America , volume 121* (issue 3), pp. 1679-1690.

[Murphy+08] Murphy, P. J., McGuigan, K., Walsh, M., & Colreavy, M. (2008). Investigation of a glottal related harmonics-to-noise ratio and spectral tilt as indicators of glottal noise in synthesized and human voice signals. *Journal of Acoustical Society of America* , 1642-1652.

[Nadeu+91] Nadeu, C., Pascual, J., & Herdondo, J. (1991). Pitch Determination Using The Cepstrum Of The One-Sided Autocorrelation Sequence. *ICASSP.*

[Nicastri+04] Nicastri, M., Chiarela, G., Gallo, L. V., Catalano, M., & Cassandro, E. (2004). Multi Dimensional Voice Program (MDVP) and Amplitud Variation Parameters in Euphonic Adult Subjects. Normative Study. *Acta Otorhinolaryngolital* , 337-341.

[Noll64] Noll, A. M. (1964). Shot-Time Spectrum and Cepstrum Techniques for Vocal-Pitch Detection. *Journal of the Acoustical Society of America , 36* (2), 296-302.

[Noll67] Noll, A. M. (1967). Cepstrum Pitch Determination. *Journal of the Acoustical Society of America , 41* (2), 293-309.

[Novorita99] Novorita, B. (1999). Incorporation of temporal masking effects into bark spectral distorsion measure. *ICASSP*, (págs. 665-668). Phoenix.

[Oleagordia+14] Oleagordia, I., & García, B. (2014) Enhancement of Shimmer and HNR in Oesophageal Speech. BioMed, 139-145. Zurich

[Oleagordia+12] Oleagordia, I., & García, B. (2012). *Patente nº BI-596-12.* España.

[Ortolan+03] Ortolan, R. L., Mori, R. N., Pereira, R. R., Cabral, C. M., Pereira, J. C., & Cliquet, A. (2003). Evaluation of Adaptive/Nonadaptive Filtering and Wavelet Transform Techniques for Noise Reduction in EMG Mobile Acquisition Equipment. *IEEE TRANSACTIONS ON NEURAL SYSTEMS AND REHABILITATION ENGINEERING , 11* (1), 60-69.

[Paliwal+87] Paliwal, K. K., & Basu, A. (1987). A Speech Enhancement Method Based on Kalman Filtering. *ICASSP*, (págs. 177 - 180).

[Paliwal+93] Paliwal, K. K., & Atal, B. S. (1993). Efficient vector quantization of LPC parameters at 24 bits/frame. *IEEE Transaction of Speech Audio Process* , 3-14.

[Park+11] Park, Y. S., & Bera, K. A. (2011). Maximum entropy autoregressive conditional heteroskedasticity model. *Journal of Econometrics* , 219-230.

[Piszczalski+79] Piszczalski, M., & Galler, B. A. (1979). Predicting musical pitch from component frequency ratios. *Journal of the Acoustical Society of America , 66* (3), 710–720.

[Proakis+07] Proakis, J. G., & Manolakis, D. G. (2007). *Digital signal processing: principles, algorithms and applications.* Prentice Hall.

[Puo04] Puo, A. (2004). Tracheoesophageal Voice restoration with total laryngectomy. *Otolaryngol Clin N Am* , 531-545.

[Qi+97] Qi, Y., & Hillman, R. E. (1997). Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals. *J. Acoust. Soc. Am. , 102* (1), 537-543.

[Qi+99] Qi, Y., Hillman, R. E., & Milstein, C. (1999). The estimation of signal-to-noise ratio in continuous speech for disorders voice. *J. Acoust. Soc. Am.* , 2532-2535.

[Qiang06] Qiang, F. (2006). Robust glottal source estimation based on joint source-filter model optimization. *IEEE Transactions on Audio, Speech and Language Processing , 14* (2), 492-501.

[Rabiner+98] Rabiner, L., & Juang, B. H. (1998). *The Digital Signal Processing Handbook.* CRC Press, IEEE Press.

[Ramírez07] Ramírez Cortés, J. H. (2007). *Tesis: Identificación de Sïstemas en Representación en Espacio de Estados.* Cuernavaca, México.

[Raymond+05] Raymond H., Colton, J. K. (October 2005). *Understanding voice problems.* Lippincott Williams & Wilkins.

[Rioul92] Rioul, O. (1992). Simple regularity criteria for subdivision schemes. *SIAM J.Math. Anal. , 23*, 1544–1576.

[Rojas99] Rojas Soriano, R. (1999). Guía para realizar investigaciones sociales.

[Rosenberg71] Rosenberg, A. (1971). Effect of glottal pulse shape on the quality of natural vowels. *J. Acoust. Soc. Amer. , 49*, 583-590.

[Ruiz+06] Ruiz, I., García, B., Vicente, J., & Méndez, A. (2006). Improvement of the shimmer of esophageal voices using Wavelet. *BioSignal*, (págs. 47-59). Brno.

[Ruiz+07a] Ruiz, I., García, B., Méndez, A., & Villanueva, V. (2007). Automated Characteritation of Esophageal and Severl injured Voices by means of Acoustic Parameters. *EUSIPCO 2007*, (págs. 2224-2228). Poznan.

[Ruiz+07b] Ruiz, I., García, B., Méndez, A., & Vilanueva, V. (2007). Oesophageal Speech Enhancement Using Kalman Filters. *ISSPIT*, (págs. 1194-1197). El Cairo.

[Ruiz+08] Ruiz, I., García, B., Méndez, A., & Mendezona, M. (2008). Oesophageal Voice Cycle Detection in Shimmer Calculation Algorithm. *ISPRA* (págs. 156-159). Cambridge: WSEAS press.

[Ruiz+09] Ruiz, I., García, B., & Méndez, A. (2009). Two Approaches of Kalman Filtering for Oesophageal Speech. *International Congress on Image and Signal Processing (CISP'09)*, (págs. 345-349). Tianjin.

[Ruiz+10a] Ruiz, I., García, B., & Méndez, Z. (2010). Two New Approaches of Kalman Filtering for Oesophageal Speech. *ISSPA*, (págs. 225-228). Kuala Lumpur, Malaysia.

[Ruiz+10b] Ruiz, I., García, B., & Méndez, A. (2010). New Approach for Oesophageal Speech Enhancement. *ISIVC 2010*, (págs. 121-124). Rabat, Marruecos.

[Ruiz+10c] Ruiz, I., Garcia, B., & Méndez, A. (2010). Using Games to Assess Oesophageal Voice. *CGAMES*, (págs. 49-54). Louisville, Kentucky, USA.

[Ruiz+12a] Ruiz, I., García, B., & Méndez, A. (2012). Using Games to Assess Oesophageal Voice. *International Journal of Science and Advanced Technology , 2* (3), 143-150.

[Ruiz+12b] Ruiz, I., & García, B. (2012). Improvement of Shimmer Parameter of Oesophageal Voices Using Wavelet Transform. En D. Baleanu, *Wavelet Transforms and their Recent Applications in Biology and Geoscience* (págs. 139-160). Croacia: Intech.

[Ruiz+12c] Ruiz, I., & Garcia, B. (2012). Enhancement of Shimmer in Oesophageal Speech. *SIIE 2012*, (págs. 194-199). Djerba, Tunisia.

[Ruiz+13] Ruiz, I., & García, B. (2013). Enhancement of Shimmer in Oesophageal Speech Using Different Wavelets. *ISSPIT.* Athens, Greece.

[Sakakibara+02] Sakakibara, K., Konishi, T., Imagawa, H., Murano, E., Kondo, K., Mumada, M., y otros. (2002). Observation of the laryngeal movements for throat singing-vibration of two pairs of the folds in humas larynx. *Acoustic Society of America.*

[Sano+89] Sano, H., & Jenkins, B. K. (1989). A neural network model for pitch perception. *Computer Music Journal , 13* (3), 41–48.

[Shannon49] Shannon, C. (1949). Communication in the presence of noise. *Proc. Institute of Radio Engineers , 37* (1), 10-21.

[Shapiro65] Shapiro, S. S. (1965). An analysis of variance test for normality (complete samples). *Biometrika* , 591-611.

[Sheng96a] Sheng, Y. (1996). *The Transforms and Applications Handbook.* Boca Raton, Florida. EE.UU.: A. D. Poularikas.

[Sheng96b] Sheng, Y. (1996). WAVELET TRANSFORM. En *The transforms and applications handbook Series* (págs. 747-827). Boca Raton, Fl (USA): CRC Press.

[Šiupšinskienė03] Šiupšinskienė, N. (2003). Quantitative analysis of professionally trained versus untrained voices. *Medicina* , 36-46.

[Sprent+00] Sprent, P., & Smeeton, N. (2000). *Applied Nonparametric Statistical Methods* (3º ed.). Chapman and Hall.

[Strang89] Strang, G. (1989). Wavelets and dilation equations:Abrief introduction. *SIAM Review , 31*, 614–627.

[Strang96] Strang, G. (1996). Eigenvalues of (↓2)H and convergence of the cascade algorithm. *IEEE Trans. Signal Processing , 44*, 233–238.

[Straub+04] Straub, D., Gefen, D., & Boudreau, M. C. (2004). *The ISWorld Quantitative, Positivist Research Methods Website*. Obtenido de http://dstraub.cis.gsu.edu:88/quant/.

[Tirado+07] Tirado, L., & Granados, M. (2007). Epidemiología y Etilogía del Cáncer de la Cabeza y el Cuello. *Cancerología* , 9-17.

[Tohidypour+10] Tohidypour, H. R., Seyyedsalehi, S. A., & Behbood, H. (2010). Comparison between Wavelet Packet Transform, Bark Wavelet & MFCC for Robust Speech Recongnition tasks. *International Conference on Industrial Mechatronics and Automation (ICIMA).* Wuhan.

[Torres08] Torres Gallardo, B., & Gimeno Gómez, F. (2008). *Anatomía de la voz* (1ª ed.).

[Toschke+08] Toschke, A. M., von, K. R., Andreas, B., & Simon, R. (2008). Risk factors for childhood obesity: shift of the entire BMI distribution vs. shift of the upper tail only in a cross sectional study. *BMC Public Health* , 100-115.

[Tse+05] Tse, D., & Viswanath, P. (2005). *Fundamentals of Wireless Communication.* Cambridge University Press.

[Vaidyanathan87] Vaidyanathan, P. (1987). Quadrature mirror filter banks, M-band extensions and perfect-reconstruction technique. *IEEE Acoust., Speech, Signal Process. Mag. , 4*, 4-20.

[Vaidyanathan+01] Vaidyanathan, P., & Vrcelj, B. (2001). Biorthogonal partners and applications. *IEEE Trans. Signal Processing , 1*, 1013–1027.

[Vazquez+05] Vazquez de la Iglesía, F., & Fernández González, S. (2005). Caracterización Acústica y Aerodinámica de la Voz esofágica. *Acta Otorrinolaringológica* , 482-487.

[Vetterli87] Vetterli, M. (1987). A theory of multirate filter banks. *IEEE Trans. Acoust. Speech Signal Processing , ASSP-35*, 356–372.

[Vetterli+95] Vetterli, M., & Kavacevic, J. (1995). *Wavelets and Subband Coding.* Prentice Hall.

[Vicente+09] Vicente, J., García, B., Méndez, A., & Ruiz, I. (2009). Audio and Image Processing Easy Learning for Engineering Students using EasyPAS Tool. En A. Z. Ashraf, *Recent Advances in Signal Processing* (págs. 27-52). Olajnija, Croacia: Intech.

[Walker31] Walker, G. (1931). On Periodicity in Series of Related Terms. *Philsophical of Transactions of Royal Society of London , 131*, 518-532.

[Wei+09] Wei, S., Hu, Y., & Wang, R.-H. (2009). A new method for mispronunciation detection using support vector machine based on pronunciation space models. *Speech Communication , 51* (10), 896-905.

[Weymuller+00] Weymuller, E., Yueh, B., Deleyiannis, F., Mphil, M., Kuntz, A., Alsarraf, R., y otros. (2000). Quality of life in head and neck cancer. *Laryngoscope* , 4-7.

[Wiener49] Wiener, N. (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series.* New York: Wiley.

[Wilcoxon45] Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics , 1*, 80-83.

[Wing-kei+95] Wing-kei, Y., Kwong-sak, & Kin-hong, W. (1995). Pitch Detection Of Speech Signal In Noisy Environment By Wavelet. *SPIE , 2491*, 604-614.

[Wong+79] Wong, D., Markel, J., & Gray, A. (1979). Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Transactions on Acoustics, Speech and Signal Processing , 27* (4), 350-355.

[Yang+02] Yang, F., Zidong, W., & Hung, Y. S. (2002). Robust Kalman Filtering for Discrete Time-Varying Uncertain Systems With Multiplicative Noises. *IEEE TRANSACTIONS ON AUTOMATIC CONTROL , 47* (7), 1179-1183.

[Yule27] Yule, G. (1927). On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Worlfer's Sunspot numbers. *Philosophical Transactions of Royal Society* , 267-298.

[Yumoto+82] Yumoto, E., & Gould, W. J. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness. *Journal of the Acoustical Society of America , 71* (6), 1544–1550.

[Zorrilla07] Zorrilla Arena, S. (2007). Introducción a la metodología de la investigación. En S. Zorrilla Arena, *Introducción a la metodología de la investigación.* Mexico Oceano: Aguilar, León y Cal.